FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Artificial Intelligence-Enhanced Activity Recognition for Vision-Based Unmanned Aerial Vehicle Detection Systems

Alam Noor



Doctoral Program in Electrical and Computer Engineering

Supervisor: Prof. Dr. Eduardo Tovar

Co-Supervisor: Prof. Dr. Pedro Alexandre Guimarães Lobo Ferreira Souto

March 16, 2026

Artificial Intelligence-Enhanced Activity Recognition for Vision-Based Unmanned Aerial Vehicle Detection Systems

Alam Noor

Doctoral Program in Electrical and Computer Engineering

Approved by:

President: Prof. Nuno Fidalgo

Referee: Prof. David Portugal

Referee: Prof. António Neves

Referee: Prof. Andry Pinto

Supervisor: Prof. Eduardo Tovar

March 16, 2026

Resumo

Veículos aéreos não tripulados (UAVs) têm sido cada vez mais utilizados em aplicações civis e comerciais devido à sua flexibilidade de utilização e à diversidade de funcionalidades, tornando-se ubíquos no espaço aéreo. Isto pode representar um problema em espaços aéreos restritos (i.e., zonas interditas a UAVs). Neste casos é necessário garantir a segurança da aviação e a eficiência da gestão do tráfego aéreo, o que exige a detecção e identificação automáticas e confiáveis de UAVs potencialmente não-autorizados.

As abordagens existentes para detecção de UAVs baseiam-se, na sua maioria, em sistemas de visão de fluxo único, utilizando imagens RGB em condições diurnas ou imagens infravermelhas (IR) em condições noturnas. No entanto, a deteção de UAVs em zonas de exclusão aérea por meio de uma estrutura híbrida de *deep learning* que integre a visão RGB–IR em dupla modalidade continua a ser um desafio. A fusão direta das modalidades RGB e IR frequentemente degrada o desempenho da detecção, resultando em localização imprecisa das caixas delimitadoras e aumento da taxa de falsos positivos. Isto deve-se ao facto da detecção baseada em RGB ser particularmente sensível a condições de baixa iluminação e clima adverso, enquanto a deteção com imagens IR apresenta dificuldades para distinguir UAVs de objetos de fundo com características térmicas semelhantes. Como consequência, o treino de modelos convencionais de *deep learning* com dados RGB–IR combinados frequentemente resulta em precisão de detecção pobre.

Noutra dimensão, o uso de câmaras de vigilância estáticas instaladas no solo oferecem cobertura limitada. Esta limitação pode ser atenuada usando sistemas de visão embarcados em UAVs legais que inspecionam as zonas de exclusão aérea a partir do ar. No entanto, câmaras embarcadas de alta resolução aumentam o consumo de energia e a latência de transmissão de vídeo, enquanto câmaras de baixa resolução reduzem o consumo energético à custa de menor precisão de detecção, especialmente de UAVs de pequeno porte. Dessa forma, equilibrar a precisão da detecção e a latência de transmissão constitui um desafio crítico para sistemas de visão baseados em UAVs.

Para enfrentar essas limitações, esta tese propõe um modelo híbrido de *deep learning* que integra representações de características RGB e IR, possibilitando a detecção precisa de UAVs tanto em condições diurnas quanto noturnas. O modelo emprega estratégias de *cosine annealing* e *rethinking transformation* para melhorar a precisão da detecção e acelerar a convergência do treino. Foram feitas experiências em ambientes externos reais conduzidos durante o dia e a noite utilizando uma câmara de vigilância estática baseada no solo, e o conjunto de dados aberto UAV-Catch foi utilizado para treino *offline*, visando aumentar a robustez do modelo. O modelo proposto alcança uma precisão média (mAP) de 65% para entradas de vídeo RGB–IR em dupla modalidade.

Para mitigar ainda mais a similaridade entre o fundo e os UAVs, bem como a variabilidade de iluminação, esta tese introduz uma Rede de Relação de *Pixels* Baseada em Grafo com Fusão Óptica (*Optical Fusion–Graph Pixel Relation Network* – OF-GPRN) para detecção precisa de UAVs

em ambientes complexos. O sistema OF-GPRN aplica fusão óptica ao nível de imagem RGB–IR para suprimir informações irrelevantes do fundo e aumentar a visibilidade do alvo. Uma Rede de Autoatenção Residual em Grafos (*Graph Residual Self-attention Network* – GRSaN) é utilizada para estabilizar a aprendizagem, melhorar a representação de características e refinar a modelação da forma do UAV. Ao construir uma representação em grafo pixel-para-nodo com uma matriz de adjacência que codifica as relações entre *pixels*, o sistema captura de forma eficaz correlações espaciais nas imagens fundidas. Resultados experimentais mostram que o OF-GPRN atinge uma perda baixa de 0,026 e uma precisão média (mAP) de 87,8% no conjunto de dados de referência UAV-Catch baseado em RGB–IR, superando detetores baseados em ResGCN existentes, que alcançam um mAP de 69,9%.

Por fim, esta tese propõe uma estrutura de *deep reinforcement learning* assistida por redes convolucionais em grafos (*Graph Convolutional Neural Network–assisted Advantage Actor–Critic* – GCN-assisted A2C) para otimizar a latência de transmissão de vídeo. A estrutura identifica regiões ótimas correlacionadas por *pixels* correspondentes a objetos (UAVs) suspeitos e transmite seletivamente essas sub-regiões em vez de imagens completas. Uma formulação dual de Lagrange combinada com *gradient descent* garante convergência estável e o cumprimento das restrições de latência. O componente GCN fornece supervisão estruturada ao modelo *Advantage Actor–Critic* (A2C), orientando a seleção das sub-regiões e melhorando a tomada de decisão. Resultados experimentais demonstram que a estrutura *GCN-assisted A2C* minimiza a latência de transmissão de *frames* de vídeo e reduz a taxa de falsas deteções em comparação com outras abordagens de *deep reinforcement learning* e modelos de estado-da-arte.

Abstract

Unmanned aerial vehicles (UAVs) are increasingly used in civil and commercial applications due to their deployment flexibility and diverse functionalities, becoming ubiquitous in airspace. This may create challenges in restricted areas, i.e., no-fly zones. In these cases, ensuring aviation safety and efficient air traffic management requires reliable automatic detection and identification of potentially intruding UAVs.

Existing UAV detection approaches primarily rely on single-stream vision systems using either RGB imagery for daytime or infrared (IR) imagery for nighttime conditions. However, detecting UAVs in no-fly zones using a hybrid deep learning framework that integrates dual-modality RGB–IR vision remains challenging. Direct fusion of RGB and IR modalities often degrades detection performance, leading to inaccurate bounding box localization and increased false positives. RGB-based detection is particularly sensitive to low-light and adverse weather conditions, while IR imagery struggles to distinguish UAVs from background objects with similar thermal characteristics. As a result, training conventional deep learning models on mixed RGB–IR data frequently yields suboptimal detection accuracy.

On another dimension, ground-based static surveillance cameras provide limited coverage. This can be improved with vision systems onboard legal UAVs that survey no-fly zones. However, high-resolution onboard cameras increase battery consumption and transmission latency, whereas low-resolution cameras reduce power usage at the cost of diminished detection accuracy, particularly for small intruding UAVs. Therefore, balancing detection accuracy and transmission latency is a critical challenge for UAV vision systems.

To address these issues, this thesis proposes a hybrid deep learning model that integrates RGB and IR feature representations to enable accurate UAV detection under both daytime and nighttime conditions. The model employs cosine annealing and a rethinking transformation strategy to enhance detection precision and accelerate training convergence. Real-world outdoor experiments were conducted during both day and night using a ground-based static surveillance camera, and the UAV-Catch open dataset was used for offline training to improve robustness. The proposed model achieves a mean average precision (mAP) of 65% on dual-modality RGB–IR video inputs.

To further mitigate background similarity and illumination variability, this study introduces an Optical Fusion–Graph Pixel Relation Network (OF-GPRN) for precise UAV detection in complex environments. The OF-GPRN system applies RGB–IR image-level optical fusion to suppress irrelevant background information and enhance target visibility. A Graph Residual Self-attention Network (GRSaN) is integrated to stabilize learning, improve feature representation, and refine UAV shape modeling. By constructing a pixel-to-node graph representation with an adjacency matrix encoding pixel relationships, the system effectively captures spatial correlations in fused images. Experimental results show that OF-GPRN achieves a low loss of 0.026 and an mAP of

87.8% on the RGB–IR UAV-Catch benchmark, outperforming existing ResGCN-based detectors with an mAP of 69.9%.

Finally, a graph convolutional neural network–assisted deep reinforcement learning framework (GCN-assisted A2C) is proposed to optimize video transmission latency. The framework identifies optimal pixel-correlated regions corresponding to suspicious objects and selectively transmits these subregions instead of entire frames. A Lagrangian dual formulation with gradient descent ensures stable convergence while enforcing latency constraints. The GCN component provides structured supervision to the Advantage Actor–Critic (A2C) model, guiding subregion selection and improving decision-making. Experimental results demonstrate that the GCN-assisted A2C framework minimizes video transmission latency and reduces false detection rates compared with other deep reinforcement learning and state-of-the-art approaches.

Acknowledgments

All praise and gratitude are due to Almighty Allah, the Most Merciful and the Most Compassionate, who granted me strength, patience, and guidance throughout this long and demanding journey. Without His blessings and will, this work would not have been possible.

This dissertation would not have reached its successful completion without the trust, guidance, and unwavering support of my supervisors, colleagues, friends, and family. My academic journey began in a small village, Hamza Abad, in District Hangu, Khyber Pakhtunkhwa, Pakistan. The path from that village to the completion of a PhD has been long and filled with challenges. There were many moments when this journey felt more like a distant dream than a reality. Rising from humble beginnings as a cow milk seller to becoming a doctoral researcher is a transformation that I still struggle to fully comprehend, and I consider it a blessing from Allah.

I owe my deepest respect and gratitude to my parents, whose sacrifices were acts of faith and love. My mother, a devoted homemaker, and my father, a driver earning a modest income, invested everything they had in my education, trusting Allah for the rest. In a society where PhD holders are rare, they enabled me not only to pursue knowledge but also to become a source of pride for my family and community. I am also profoundly grateful to my extended family: my uncles, Shahzad Gul and Alif Khan; my brother, Muhammad Rasool; and my grandmother, whose prayers and insistence on education laid the foundation for my academic aspirations.

My heartfelt gratitude goes to my wife, Iqra, whose constant love, patience, prayers, encouragement, and sense of humor sustained me throughout this demanding journey. I am equally thankful to my son, Muhammad Jawad Noor, and my daughter, Fatima Noor, whose presence gave meaning to every sacrifice made along the way. Their unconditional support was a source of spiritual strength during difficult times. Without my family's unwavering support and prayers, this achievement would not have been possible. This thesis is dedicated to all of them.

I would like to express my sincere and profound appreciation to my supervisor, Professor Eduardo Tovar, for his continuous guidance, encouragement, and trust throughout my PhD journey. I believe Allah placed exceptional mentors in my path, and Professor Tovar has been one of the greatest blessings of this journey. His intellectual rigor, dedication to excellence, and passion for research shaped me into a better researcher and academic. Engaging in thought-provoking discussions and exploring emerging research areas under his mentorship has been both inspiring and transformative. As Director of CISTER, he created a truly international research environment and continuously supported me academically and financially through research grants and funding, enabling me to complete my PhD smoothly over the past five years.

I also extend my sincere thanks to my former supervisor, Dr. Kai Li, for having introduced me to the topic of this thesis and for his permanent quest for scientific excellence, and to my co-supervisor, Professor Pedro Alexandre Guimarães Lobo Ferreira Souto, for his invaluable support,

particularly during the thesis submission process. His professionalism, humility, and kindness reflect not only his excellence as a scientist but also his integrity as a human being.

My sincere appreciation goes to Professor Luís Almeida, my research project funding supervisor and Director of the Doctoral Program in Electrical and Computer Engineering. I am deeply grateful for his support over the past two years, especially during critical moments involving urgent grant applications. His timely assistance, wisdom, and generosity were instrumental in the completion of this work. I also wish to express special thanks to Professor Almeida for his expertise and substantial contributions to the final part of this dissertation.

I had the privilege of collaborating with many outstanding researchers and colleagues across different institutions, including (but not limited to) Luís Almeida, Anis Koubâa, Muhammad Daudi, Murray J. Corke, Adel Ammar, and Rahim Khan. I am thankful to Allah for surrounding me with such knowledgeable and supportive individuals. This dissertation would not exist without their valuable discussions, guidance, and encouragement.

I am also grateful to my colleagues and friends at CISTER for their companionship and countless enriching discussions, especially Harrison, Jing, Abdul, Saeid, Ishfaq, Miguel, Patrick, Mubarak, José, Konstantinos, David, Yimin, Shardul, and Yousef. I sincerely thank Sandra, Inês, Cristiana, Marwin, and Benilde for their tireless and invaluable administrative support, which greatly eased my academic journey.

This work was supported by the CISTER Research Unit (UIDP/UIDB/04234/2020) and the project ADANET (PTDC/EEICOM/3362/2021), funded by National Funds through FCT/MCTES (Portuguese Foundation for Science and Technology). This dissertation also benefited from the project NORTE-01-0145-FEDER-000062 (RETINA), supported by the Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF), and LASI LA/P/0104/2020.

Finally, I pray that Allah accepts this effort, forgives my shortcomings, and allows this work to be a source of benefit for humanity. Ameen.

Alam Noor

"Indeed, the patient will be given their reward without measure"

Qur'an 39:10

Contents

1 Introduction **1**
- 1.1 Motivations and Challenges . . . 4
  - 1.1.1 Motivations . . . 4
  - 1.1.2 Challenges . . . 5
- 1.2 Overview of Contributions . . . 6
  - 1.2.1 A Hybrid Deep Learning Dual Visions System . . . 6
  - 1.2.2 Fusion Flow-enhanced Graph Pooling Residual Networks . . . 7
  - 1.2.3 Latency Efficient Edge Vision System . . . 7
- 1.3 Dissertation Structure . . . 8

2 Background and Related Work **10**
- 2.1 Hybrid Deep Learning Background Study and Limitations . . . 11
- 2.2 Fusion Flow-enhanced Graph Pooling Residual Networks Background Study and Limitations . . . 16
- 2.3 Latency Efficient Edge Vision System Background Study and Limitations . . . 18
- 2.4 Technical Pillars . . . 20
  - 2.4.1 Feature Pyramid Network . . . 20
  - 2.4.2 Graph Neural Networks . . . 22
  - 2.4.3 Reinforcement Learning . . . 23
- 2.5 Chapter Summary and Concluding Remarks . . . 24

3 A Hybrid Deep Learning Model for UAVs Detection in Day and Night Dual Visions **25**
- 3.1 Hybrid Deep Learning System Model . . . 27
  - 3.1.1 Dataset Preparation and Preprocessing . . . 27
  - 3.1.2 Backbone of the UAV detection . . . 28
  - 3.1.3 Dual-stream video processing . . . 30
  - 3.1.4 Photometric and Geometric Augmentation . . . 31
  - 3.1.5 Cosine Annealing . . . 32
  - 3.1.6 Focal Loss . . . 33
- 3.2 Experiments and Results . . . 33
  - 3.2.1 Results . . . 34
- 3.3 Chapter Summary and Concluding Remarks . . . 38

4 Fusion Flow-enhanced Graph Pooling Residual Networks for UAVs Surveillance in Day and Night Dual Visions **40**
- 4.1 The Proposed OF-GPRN System Model . . . 42
  - 4.1.1 RGB-IR Fusion . . . 43

4.1.2 Graph Residual Split Attention Network (GRSaN) . . . 44
4.1.3 Graph-Pooling Feature Pyramid Network Mapping . . . 46
4.1.4 Loss Function . . . 48
4.2 Experiments and Performance Analysis . . . 49
4.2.1 Experimental Training . . . 49
4.2.2 Datasets Pre-processing . . . 50
4.2.3 Effect of Optical Fusion . . . 51
4.2.4 Effect of Region Adjacency Graphs . . . 51
4.2.5 Residual Split Attention Network Effect . . . 55
4.3 Chapter Summary and Concluding Remarks . . . 59

5 Latency Efficient Edge Vision Based System for UAV2UAV Detection 60
5.1 Mathematical System Modeling . . . 62
5.1.1 Accuracy Constraints Strategy . . . 63
5.1.2 Problem Formulation . . . 64
5.1.3 Complexity . . . 65
5.2 Proposed GCN-Assisted DRL System Model . . . 65
5.2.1 CMDP Formulation and Information Flow . . . 66
5.2.2 GCN-A2C for Decision Making . . . 66
5.2.3 Hidden Pixels Relation Exploration . . . 67
5.2.4 A2C for T_f Optimization with $\mathscr{A}_s$ Satisfaction . . . 68
5.3 Experimental Work . . . 70
5.3.1 Implementation of GCN-assisted A2C Model . . . 70
5.3.2 Ablation Study . . . 71
5.3.3 State-of-the-art (SOTA) comparisons . . . 73
5.3.4 Evaluation Results . . . 74
5.4 Chapter Summary and Concluding Remarks . . . 77

6 Conclusion and Future Work 78
6.1 Dissertation Conclusion . . . 78
6.2 Future Work . . . 79
6.2.1 Dual Day Night Vision Systems . . . 80
6.2.2 Environmental and Background Effects in Vision Systems . . . 80
6.2.3 Transmission Latency Optimization . . . 81

References 83

List of Figures

1.1 Framework of UAVs in critical areas. 1
1.2 Small UAV incidents in different no-fly zone areas caused severe damage to both the economy and lives. 2
1.3 Unauthorized UAVs patrol in the no-fly zone, which disturbs the takeoff and landing of the flight or spies on confidential military operations. 3

2.1 EdgeDuet non-optimal region selection for the transmission of small objects to server. 20

3.1 The proposed architecture for the UAV detection using dual streams. 26
3.2 The first row is of RGB video frames, and the second row is of IR video frames from different videos and background variations of different sizes of UAVs. . . . 27
3.3 Green bounding boxes, such as additional boxes, missing boxes, and shifted boxes, indicate anti-UAV dataset annotation errors. While red bounding boxes show the correct version. 27
3.4 The compound scaling of width, depth, and resolution uniformly in a fixed ratio. . 30
3.5 A top-down and bottom-up bidirectional approach to feature fusion, ensuring accurate object classification and bounding box prediction. 31
3.6 Photometric and geometric augmentation of UAV with a boundary box for precise bounding box localization. 32
3.7 Classification and regression Loss using dual stream with the augmented fast convergence and non-augmented classification and regression loss has been shown in this figure. 35
3.8 In this figure the batch classification (BC) and regression (BR) losses for both augmentation and non-augmentation using boundary box geometric augmentation techniques are shown for the dual stream training. 35
3.9 Mean average precision (mAP) from the training for both streams with and without photometric and geometric augmentation and bounding box augmentation. 36
3.10 Training and validation loss of the hybrid deep learning model with dual stream UAV detection using 100 epochs with and without photometric and geometric augmentation. 36
3.11 Validation classification (VC) and regression (VR) losses of proposed model for both augmented data and non-augmented boundary boxes are shown in this figure. 37
3.12 UAV detection from an RGB sample of day vision from test data. 37
3.13 UAV detection from infrared (IR) videos of night vision samples from test data. . 37
3.14 UAV detection from real-time videos of on-campus samples from test data. . . . 38

3.15 Miss-classification samples of UAV from different testing videos of RGB, IR videos of both day-night vision of the UAV Catch dataset, and on-campus testing. 38

4.1 The proposed OF-GPRN system architecture for UAV detection and tracking in day and night vision. 42
4.2 In a deep overview of the split graph attention is shown here, which represents each layer of the graph convolutional network. 47
4.3 The scaling factor's impact on cross-entropy to automatically downweight loose samples during training. 48
4.4 The columns on the far right are RGB frames with object optical flow. The middle column shows the IR frames and optical flow. In contrast, the most right-hand column combines RGB-IR and then displays the effect of UAV optical flow. . . . 52
4.5 The region adjacency graphs before optical flow and after optical flow have three different segment approaches: The leftmost column is the Felzenszwalb algorithm performance, the middle column is the SLIC algorithm results, and the rightmost column is the Quickshift algorithm segmentation for the adjacency matrix. 53
4.6 The training losses of the OF-GPRN (based on the Split Attention Network) with three different segment approaches for region adjacency graph algorithms after fusion plus optical flow. 54
4.7 The training loss of the four different graph residual networks is: The cyan line represents the ResGCN with RGB-IR, the green line shows the fusion-based ResGCN, the blue line represents the optical flow fusion ResGCN, and the red line represents the proposed optical flow-assisted graph pooling residual network (OF-GPRN). Moreover, (a) is trained with a learning rate of 10^{-4} and (b) is trained with a learning rate of 10^{-6}. 55
4.8 The mean average precision (mAP) performance of the proposed OF-GPRN model vs. the RGBIR ResGCN, Fusion ResGCN, and optical flow ResGCN. Moreover, (a) is trained with a learning rate of 10^{-4} and (b) is trained with a learning rate of 10^{-6}. 56
4.9 The proposed OF-GPRN model performance in the form of training, validation, and testing losses uses Quickshift for the region adjacency graphs. Moreover, advanced focal loss is used to train the proposed model. 56

5.1 On-board UAV processing video to detect suspicious UAVs. 60
5.2 The figure illustrates the proposed framework, indicated by several color-coded parts. The left-side black box encloses the proposed model region, which symbolizes the core framework of the GCN-assisted A2C. The blue box shows the actor who selected the action. While the critic's role in action-value evaluations. The green box is the GCN future state prediction, which regulates actor behavior to optimize latency and minimize errors. 67
5.3 The ablation study of different RL algorithms performance using loss using data that have across different sizes of UAVs and dynamic movements with different environmental effects. 72
5.4 With the proposed GNN-A2C (GCN-assisted A2C) and different DRL algorithms, we can check the average (30FPS) IoU and tracking AP accuracy. 73
5.5 The effectiveness of GCN-assisted A2C over SOTA models to optimize the transmission latency. 74

5.6 Performance of different models under UAV-to-server video sub-area of frame transmission. In (a), the AP and meanIoU accuracy of finding suspicious UAVs from end-to-end are shown. In (b), the meanIoU over end-to-end latency is compared. And in (c), the performance of GCN-assisted A2C is shown over different input video frame resolutions. 75
5.7 We evaluated three different sizes of UAVs—small (S), medium (M), and large (L)—under various conditions, such as clear sky, complex background, and partially cloudy. 76

List of Tables

2.1 Searching Procedure with Different Research Databases and Keywords Used . . 10
2.2 Comprehensive Literature Review on UAV Detection 13

3.1 UAV dataset statistical approach of the RGB and IR video frames. 28

4.1 Basic Notations 42
4.2 Experimental Results with Stability and Model Performance 50
4.3 Summary of Segmentation Methods 54
4.4 Performance Comparison of Detection Models with GPUs 58

Chapter 1

Introduction

Unmanned Aerial Vehicles (UAVs) are commonly used for commercial purposes due to flexible deployment and versatility, as shown in Figure 1.1. Examples of applications include public surveillance, cartography, search and rescue, data collection, data aggregators, and task uploading using artificial intelligence algorithms and Internet of Things strategies (Li et al., 2022c, 2023, 2022b), road maintenance inspection (Zhu et al., 2024), monitoring urban and rural objects (pedestrians, vehicles, bicycles, etc.) (Ning et al., 2024) (Jiang et al., 2024) (Li et al., 2024b) (Fan et al., 2025), and vegetation growth and health monitoring (Laghari et al., 2024). The global market size of UAVs is growing and is expected to reach 91.23 billion US dollars by 2030 (Insights, 2024).

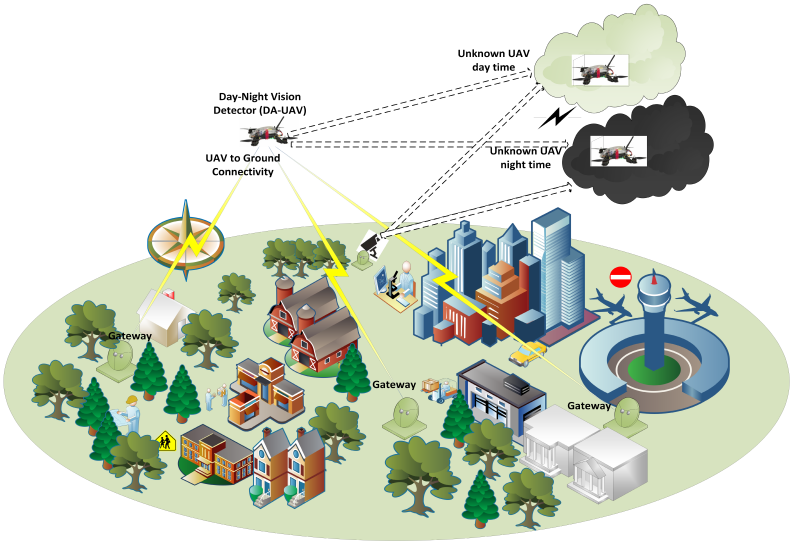


Figure 1.1: Framework of UAVs in critical areas.

However, this wide use of UAVs also brought threats to the environment and the public as well, a result of their easy usage even in forbidden areas. There are several cases where small UAVs challenged aviation safety and interrupted efficient air traffic control in no-fly zone areas (POST, 2020; Shackle, 2020; Hradecky, 2017; Latvian Public Broadcasting LSM, 2025, 2020). In one of

them a single drone entered Gatwick airport shortly before Christmas and closed the runway for 33 hours, which costed over 50 million Pounds. According to the airport announcement, around 1000 flights were cancelled, which not only caused economic loss but also affected thousands of passengers. Similarly, a small UAV crashed into a Boeing 737 jet plane during landing at an airport in Mozambique, but luckily no passenger was injured.

Small UAVs are not only a challenge for airports airspace but also for playgrounds, public arenas, chemical plants, industries, and laboratories, as shown in Figure 1.2. Moreover, flying UAVs in big cities is rather dangerous since they can affect crowds, urban life and traffic, and can create different difficulties.

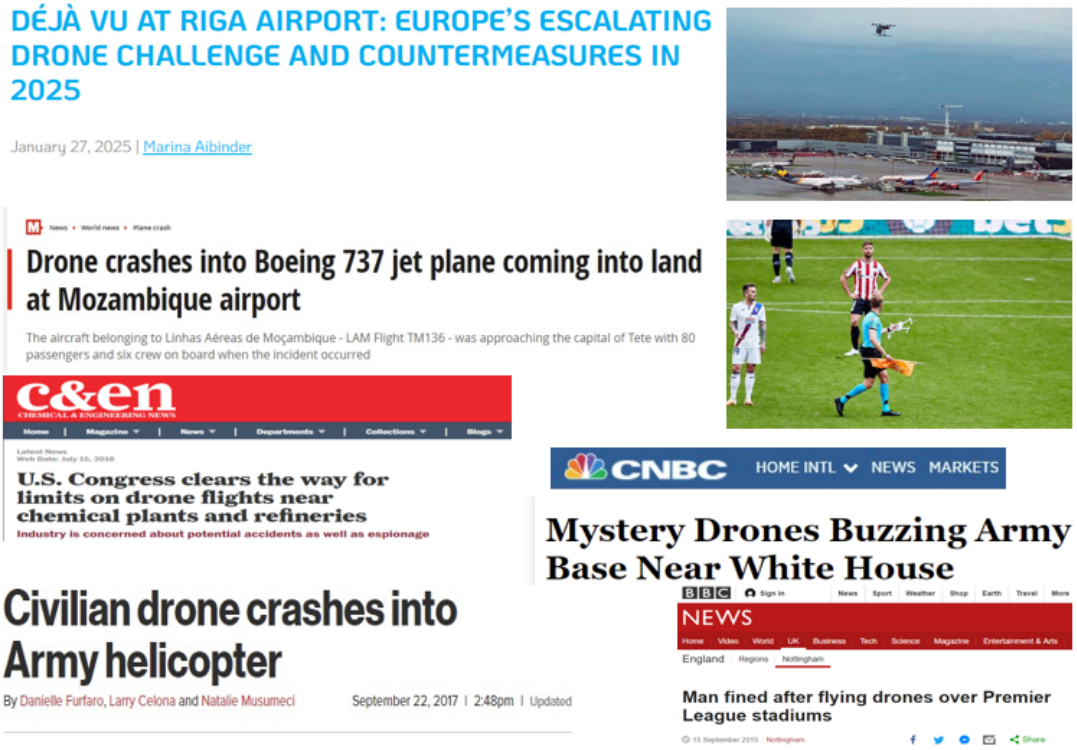


Figure 1.2: Small UAV incidents in different no-fly zone areas caused severe damage to both the economy and lives.

Unauthorized UAVs also pose a risk to sensitive infrastructure, such as power plants and communication networks (Li et al., 2019; Kurunathan et al., 2024), where an accidental or intentional collision could result in widespread service disruptions or catastrophic failures (Jiang et al., 2020). As shown in Figure 1.3, an unauthorized UAV outfitted with cameras or other surveillance devices in a no-fly zone can infringe on privacy rights and present substantial security threats to both civilian and military operations. Such UAVs have the potential to obtain unauthorized imagery or data, thereby providing malicious entities with invaluable information (Akter et al., 2022).

Therefore, we need an overall detection system with integrated communication and perception

that performs effectively in both day and night illumination alongside complex backgrounds, including birds, buildings, and cloud environments, and that works well even when bandwidth and computing power are limited. The goal is not only to identify, but also to localize and track in real time so that human operators or self-driving systems can take timely action to mitigate the danger and prevent damage.

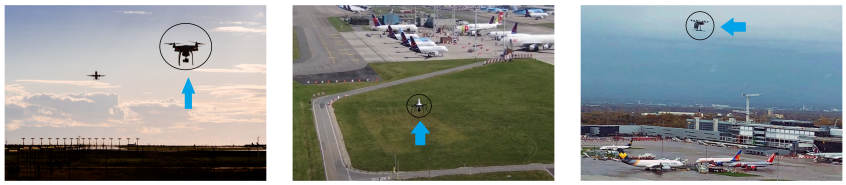


Figure 1.3: Unauthorized UAVs patrol in the no-fly zone, which disturbs the takeoff and landing of the flight or spies on confidential military operations.

Unfortunately, detecting small UAVs cannot be done by current advanced radar systems that either do not detect them or generate false positive detections. Moreover, radar systems also fail in a number of situations, such as building corners, objects near the ground or reflective surfaces, and difficult weather conditions.

One essential feature of such a detection system is the ability to distinguish a UAV from objects with similar shape, such as birds or aircraft. In addition to detecting limitations, camera-based systems also have unique challenges, such as motion blur at long camera ranges, a small target size that appears to be present, and a lot of similarity between the foreground and background.

This dissertation primarily examined these practical and theoretical core technical challenges and limitations. UAV detection has been a long-standing, essential, and complex scientific topic for a decade, and the public and military sectors have seen an increase in UAV monitoring for different purposes and protection. Artificial intelligence (AI) vision systems are essential and have the capabilities to find an object in the air with very high precision and to track it for a long time.

Researchers created AI models that use radar-based noise signals and radiofrequency (RF) signal methods to find UAVs (Sazdić-Jotić et al., 2022; Basak et al., 2022, 2021; Stoica et al., 2020; Allahham et al., 2020). These models either use moving components like motors, propellers, and chassis vibration to detect UAVs or record video directly and divide it into multiple frames for deep neural network (DNN) methods. However, using an AI model for radar-based UAV detection to protect no-fly zones is not practical due to the radar limitations with small UAVs as referred before. Similarly, automatic background removal techniques and algorithms are often used in image-based detection systems (Saqib et al., 2017; Seidaliyeva et al., 2020a), but background subtraction is not effective when the object and background have high similarity. In such instances, a hybrid system that combines complementary modalities (RGB and IR) and motion information (optical flow) appears more promising, especially when combined with models that accurately demonstrate how regions and proposals in the scene connect to each other.

Emerging developments in AI provide new ways to address these problems, particularly feature pyramids that can be used for multi-scale robustness, graph neural networks that can be used for contextual reasoning, and reinforcement learning that can be used for variable data transmission policies. We studied these novel ideas and applied them to a dual (RGB/IR) day–night vision system in real-world operations with limited edge space.

The remaining part of this chapter provides a brief description of the motivation, the challenges addressed, and the scientific contributions emerging from this work. In Section 1.1, we explain the motivations and challenges of this dissertation. In Section 1.2, a summary of the most significant innovations is described. While in the final section of this dissertation, Section 1.3, the organization of the following thesis structure is described. To be fully transparent, we provide a brief overview of the evaluation setup (datasets, metrics, and ablations) to assist the reader in understanding what to expect in the following chapters.

1.1 Motivations and Challenges

1.1.1 Motivations

Advanced models for UAV detection needs AI vision detection and classification systems for public safety to address the security issues posed by the rapid proliferation of UAVs, mainly due to their long-term dependability, inexpensive cost, simplicity of installation, and adaptability (Behera and Bazil Raj, 2020). The AI vision system models seek to identify instances of small UAVs from various categories within an image or video and provide the spatial position and extent of each occurrence. Moreover, these AI vision-based UAV detections might bring positive advancements and reduce the cost of task processing time with high accuracy for humans. There are plenty of attempts at UAV detection using single-stream (day or night vision) AI vision system models. Most of these models are based on deep learning models are developed for UAV detection based on either RGB (the day vision camera) or IR (the night vision one) (Rozantsev et al., 2017; Saqib et al., 2017; Dong and Zou, 2017; Peng et al., 2018; Nalamati et al., 2019), while other deep learning models have been utilized for UAV detection based on their color congruence with the background. Notably, these models often demonstrate proficiency in detecting UAVs in daytime color images or in infrared (IR) images captured at night (Li et al., 2021b; Patel et al., 2023; Ashraf et al., 2021; Isaac-Medina et al., 2021b).

Real deployments must operate across all times and weather, where illumination changes, motion blur, and scale variation degrade performance. This motivates our hybrid and graph-enhanced approach that is explicitly designed for dual-vision inputs and small-object saliency. In addition, practical systems must obey latency and energy limits on UAV/edge devices and cope with variable network conditions during live streaming. So there are some challenges that need to be addressed to achieve UAV detection in any situation, from light variations to high background-UAV similarities, speed, and size of the UAVs. To address these challenges, we target the following main and key research topics.

1.1.2 Challenges

- **Challenge 1: Hybrid Model for Day-Night Vision**. As we studied in the previous section, most deep learning models are developed for the UAV detection based on either RGB (the day vision camera) or IR (the night vision one) (Rozantsev et al., 2017; Saqib et al., 2017; Dong and Zou, 2017; Peng et al., 2018; Nalamati et al., 2019).. However, the UAV detection rate based on RGB is low when the camera has insufficient light in the daytime, e.g., cloudy or stormy weather. The UAV detection based on IR is affected when the UAV overlaps with an object (e.g., buildings or trees) in the background. Due to the effect of the light condition or similar shape of the UAV and other objects, false detection of the UAV and results in a low training accuracy of the deep learning with RGB or IR. Developing a hybrid deep learning model for processing RGB and IR videos is non-trivial since the RGB and IR video frames are trained independently with day or night vision features. As a result, the training for RGB and IR videos can not be directly combined for the UAV detection in a dual vision mode. Moreover, developing a hybrid model may introduce additional complexity due to cross-modal feature mismatches when RGB and infrared videos are jointly used for training, which can complicate feature fusion in deep learning models (Zhang et al., 2019). Similarly, the AI models that were made for IR night vision had problems with changes in light levels or when the sun went down or came up. It is important to have a unique hybrid model that should work for both day and night visions. Moreover, the hybrid model should preserve the features for both RGB and IR videos during training with both types of data. The hybrid model must tackle all types of day and night visibility and needs to distinguish between UAVs and other objects. Our formulation addresses these points by (i) multi-scale fusion (FPN/BiFPN), (ii) balanced sampling across modalities, and (iii) training schedules (cosine annealing, photometric and geometric augmentations) that stabilize convergence.

- **Challenge 2: Background and UAV similarities**. Deep learning models have been utilized for UAV detection based on their color congruence with the background (Li et al., 2021b; Ashraf et al., 2021; Isaac-Medina et al., 2021b). Notably, these models often demonstrate proficiency in detecting UAVs in daytime color images or in infrared (IR) images captured at night. However, its efficacy tends to diminish when faced with homogenous backgrounds at nighttime or fluctuating illumination during daylight hours. The UAV detection rate based on RGB is low when the camera has insufficient light in the daytime, e.g., cloudy or stormy weather. The UAV detection based on IR is affected when the UAV overlaps with an object (e.g., buildings or trees) in the background. Due to the effect of the light condition or similar shape of the UAV and other objects, false detection of the UAV results in low training accuracy for deep learning with RGB or IR. It is important to diminish or remove the static background and lighting when faced with homogenous backgrounds at nighttime or fluctuating illumination during daylight hours. We therefore combine optical flow with graph-based residual pooling (OF-GPRN) to suppress static background and emphasize motion-consistent regions, improving small-object AP without losing fine details.

- **Challenge 3: On-Board UAV Latency Optimization** Ground cameras cover a limited area to detect the suspicious UAV. It is required to have an on-board UAV system to detect suspicious UAVs. However, on-board UAV systems are resource-constrained by limited battery power. Using a high-resolution on-board camera consumes a large amount of battery power and also extends the video transmission latency during sending video frames from UAV to edge server. While low-resolution cameras save battery power, however, they create problems recognizing small UAVs during different environmental conditions. Moreover, the detection of the on-board system model is very low. Moreover, mobile edge computing (MEC) as UAVs is used for a wide spectrum of mobile vision applications, such as augmented reality (Hitchcock and Sung, 2018), trajectory planning (Li et al., 2023), or visual surveillance systems (to prevent problems in no-fly zones) (Noor et al., 2021; Ullah et al., 2023; Rezaee et al., 2022). However, the system should be accurate, real-time, and resource-efficient. To perform real-time tasks in computer vision, the UAVs should be capable of detecting large and small objects in dynamic environments with high accuracy. Moreover, UAVs are mostly used as MECs to stream videos to a server and get important data from the server after preprocessing to identify the objects (Galanopoulos et al., 2021; Xu et al., 2023). However, real-time video transmission and receiving tasks between UAVs and servers require a longer delay to transmit video frames and achieve high accuracy. As a result, neither their execution on resource-constrained user devices such as UAVs nor their transfer to remote cloud servers is practicable because of higher transmission delay and low accuracy (Dong et al., 2022). We cast this as minimizing the balancing latency $T_{t,n}$ and accuracy $\mathscr{A}_s$ constraint. Our GCN-assisted A2C policy selects pixel-correlated regions to transmit, reducing bandwidth while improving detection quality.

1.2 Overview of Contributions

This section presents the primary contributions of this dissertation addressing the research challenges defined in the previous section 1.1.2. The section is organized within three fundamental topics: (1) a hybrid deep learning model for detecting UAVs; (2) a network that uses optical flow to help with graph pooling for UAV detection; and (3) a method that uses GCN to improve deep reinforcement learning for reducing transmission delays. Each contribution is experimentally validated with ablation studies and comparisons to state-of-the-art baselines, and we report both accuracy and latency metrics and system model metrics.

1.2.1 A Hybrid Deep Learning Dual Visions System

In this section, we address **challenge 1**. We propose a new hybrid deep learning model to detect the UAVs in day and night visions with a high detection precision and accurate bounding box localization. The proposed hybrid deep learning model is developed with cosine annealing and photometric and geometric augmentation to improve the detection precision and accelerate the training convergence. To validate the hybrid deep learning model, real-world experiments are

conducted outdoors in daytime and nighttime, where a surveillance video camera on the ground is set up for capturing the UAV. In addition, the UAV-Catch open database is adopted for offline training of the proposed hybrid model, which enriches training datasets and improves the detection precision. The experimental results show that the proposed hybrid deep learning model achieves 65% in terms of the mean average detection precision given the input videos in day and night visions. The previously mentioned contributions are elaborated upon in the following publication:

- **A. Noor**, K. Li, A. Ammar, A. Koubaa, B. Benjdira and E. Tovar, "A Hybrid Deep Learning Model for UAVs Detection in Day and Night Dual Visions", 2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI), Atlanta, GA, USA, 2021, pp. 221-231, doi: 10.1109/CogMI52975.2021.00037.

1.2.2 Fusion Flow-enhanced Graph Pooling Residual Networks

In this section, we address **challenge 2**. We propose a new optical flow-assisted graph-pooling residual network (OF-GPRN), which significantly enhances the UAV detection rate in day and night dual visions. The proposed OF-GPRN develops a new optical fusion to remove superfluous backgrounds, which improves RGB/IR imaging clarity. Furthermore, OF-GPRN extends optical fusion by incorporating a graph residual split attention network and a feature pyramid, which refines the perception of UAVs, leading to a higher success rate in UAV detection. A comprehensive performance evaluation is conducted using a benchmark UAV catch dataset. The results indicate that the proposed OF-GPRN elevates the UAV mean average precision (mAP) detection rate to 87.8%, marking a 17.9% advancement compared to the residual graph neural network (ResGCN)-based approach. The previously mentioned contributions are elaborated upon in the following publication:

- **A. Noor**, K. Li, E. Tovar, P. Zhang, B. Wei, "Fusion flow-enhanced graph pooling residual networks for Unmanned Aerial Vehicles surveillance in day and night dual visions", Engineering Applications of Artificial Intelligence, Volume 136, Part B, 2024, 108959, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2024.108959.

1.2.3 Latency Efficient Edge Vision System

In this section, we address **challenge 3**. We propose a graph convolutional neural network-assisted deep reinforcement learning system model (GCN-Assisted A2C) to find the optimal pixel-correlated area of a suspicious object. While we use Lagrangian dual form with gradient descent for the proposed system model to prevent lack of convergence and over- and under-penalization constraint violation during latency optimization. Moreover, the proposed system model sends a sub-group pixel-correlated area of the frame from the UAV to the server rather than the transmission of the whole frame. The proposed framework utilizes the GCN model to explore hidden representations of feature-correlated groups of pixels. Moreover, GCN supervises the A2C model, which selects a subgroup to enhance the transmission latency, thereby supervising the training

of UAV actions in A2C. The experimental results show that GCN-Assisted A2C achieves minimum video frame transmission latency and reduces the risk of false detection rate in UAV vision systems over other DRL and state-of-the-art models. The previously mentioned contributions are elaborated upon in the following publication that is under submission:

- **A. Noor**, L. Almeida, K. Li, J. Wu, E. Tovar, "Graph Neural Assisted Actor-Critic for Latency-Efficient Edge Vision System", (under review).

1.3 Dissertation Structure

We organized the remaining dissertation as follows.

Chapter 2 explores the background and related work of this dissertation to understand why we conducted the study in the UAV detection field under different scenarios. First, the important thing is to define the research databases and keywords needed to explore the research gap. We explored different research databases like Elsevier, ACM/IEEE, Nature, Springer, Wiley, MDPI, and other EI journals and top conferences. For UAV, we also include the keyword "drone" and search for all related secondary key terms relevant to this dissertation. Next, we provide the overview of dual vision, background and foreground similarities, and transmission latency optimization. We present the rationale behind the application of artificial intelligence algorithms in our proposed research. Moreover, in different sections, we analyzed state-of-the-art works related to day and night object detection, background and foreground similarities, and transmission latency optimization. The limitations of these studies are given in a single and concise paragraph for each section, and it is highlighted why our proposed work is better than related works. Finally, we concluded with the summary of the chapter and connection with other chapters. We also include comparison tables to help the reader quickly understand trade-offs among detection methods of different models.

Chapter 3 details the hybrid deep learning model as a single model for day and night dual visions. It also presents template matching for the correction of temporal annotation and UAV localization. Additionally, the chapter provides a detailed presentation of the bidirectional feature pyramid network, photometric and geometric augmentation, and cosine annealing. The chapter shows numerous experimental results of the proposed hybrid deep learning model and a comparison with the state-of-the-art EfficientDet model using CVPR Anti-UAV Benchmark database. Implementation details (training schedule, augmentation, and hardware) are provided for reproducibility.

Chapter 4 studies the background and UAV similarities in both RGB and IR videos and proposes a novel Fusion Flow-enhanced Graph Pooling Residual Networks system model. The chapter explains the RGB-IR Fusion (and its development) by separating the input frames into base layers, fine structure, and coarse structure; background subtractions using optical flow; graph residual split attention network; graph to pooling feature pyramid network mapping; and consistent cross-entropy loss function. The performance of the model is evaluated by using the Anti-UAV Benchmark database. Moreover, experimental work studies the effect of optical fusion, the effect

of region adjacency graphs, and consistent cross-entropy loss function performance with SLIC, Felzenszwalb, and Quickshift superpixel segmentation techniques. The chapter ablation study was conducted with optical flow ResGCN, Fusion ResGCN, RGBIR ResGCN, and the proposed OF-GPRN. The chapter shows numerous experimental results comparing the proposed OF-GPRN system model with the state-of-the-art models. We also examine the failure circumstances (e.g., rapid blurred edges and significant distortions) in order to inform future enhancements.

Chapter 5 presents a novel graph convolutional network-assisted deep reinforcement learning model (GCN-Assisted-A2C) for transmission latency optimization and UAV-to-edge server efficiency. The chapter presents the mathematical system modeling and effect of the frame area of the transmission latency under the accuracy constraints. The chapter conducts a comprehensive assessment of the proposed GCN-Assisted-A2C using public datasets on Roboflow from YouTube. The chapter on the implementation and ablation phase studied A2C, DDPG, and DQN with a comparison of the proposed GCN-Assisted-A2C system model for each frame of videos. Experimental results demonstrate that GCN-Assisted-A2C offers superior transmission latency minimization capability compared to the state-of-the-art methods of FlexPatch and EdgeDuet.

Chapter 6 presents a summary of the dissertation. This chapter concluded by addressing the limitations of the dissertation and highlighting possibilities for future research. In particular, we outline extensions to multi-UAV cooperation, event-camera integration, and privacy-preserving region selection.

Chapter 2

Background and Related Work

This chapter surveys the technical background and related work required to understand our contributions across three dimensions: (i) hybrid day–night vision for UAV detection, (ii) fusion flow–enhanced graph neural models for robust small-object detection, and (iii) latency-efficient edge vision systems for live video. We first outline the scope and search strategy, then review prior work and its limitations for each dimension, and finally distill the gaps that motivate our approach. In Section 2.1 we review object detection in RGB/IR and hybrid settings and position our hybrid model. Section 2.2 analyzes state-of-the-art (SotA) detection methods most relevant to our Fusion Flow–enhanced Graph Neural Network. Section 2.3 reviews latency-efficient edge vision and how our system differs, particularly for transmission latency optimization. Our literature search categories and databases are summarized in Table 2.1.

Table 2.1: Searching Procedure with Different Research Databases and Keywords Used

Searching Procedure		
Category	**Primary Search**	**Secondary Search**
Search Databases	Scopus	Elsevier, ACM/IEEE, MDPI
	Science Direct	Nature, Springer, Wiley
	Google Scholar	All
	Engineering Index	All EI journals and conferences

(continued on next page)

Searching Procedure (continued)		
Category	**Primary Search**	**Secondary Search**
Search Keywords	UAV detection	UAV to UAV, UAV real-time detection, RGB videos for UAV, IR videos for UAV, Deep learning-based UAV detection, Deep learning-based UAV classification
	UAV classification	UAV and other object classification, different size UAV classification
	Unmanned Aerial Vehicles	Unmanned Aerial Vehicles detection, Deep learning-based Unmanned Aerial Vehicles detection
	Aircraft detection	Aircraft detection, Deep learning-based aircraft detection
	Drone detection	Drone detection, Deep learning-based drone detection
	Drone classification	Drone classification, Deep learning-based drone classification
	Counter UAV	Deep learning-based counter UAV

2.1 Hybrid Deep Learning Background Study and Limitations

Classical and learning-based detectors have been developed on either RGB or IR imagery to detect objects such as people, vehicles, and UAVs. Single-stream pipelines typically exploit motion/spectral cues or learned appearance features; however, they struggle across day–night conditions and background similarity. Below, we summarize representative approaches and their constraints.

Several detection approaches with RGB or IR are studied in the literature to detect objects, such as people or vehicles. Molchanov et al. (2013) and Baek et al. (2017) presented micro-Doppler signatures; thermal position intensity histograms of the oriented gradient are adopted for detecting one stream. Wang et al. (2010) used aggregated channel features and shape context descriptor-based pedestrian detection are presented for the detection in night environments. Guan et al. (2019) combines extracted features of two imaging sensors by using illumination and coefficient correction of the day and night estimated illumination. Many algorithms in the literature are used for UAV detection; infrared-based systems, such as background subtraction, thermal visible video fusion, and robust multi-stage approaches, detect objects with cameras (Davis and Sharma, 2007; Leykin et al., 2007; Chen et al., 2012). Shi et al. (2018) used Hidden Markov models to extract features and detect UAVs in noisy environments. Moreover, Li et al. (2012) used a histogram of an oriented gradient to adjust parameters and geometric characteristics and a support vector machine for the object detection in thermal images. Teutsch et al. (2014) presented a two-stage person recognition model that adopts maximally stable extremal regions and discrete cosine transform with random naïve Bayes. Video-based object detection models are developed to detect objects based on extracting discriminant features such as distinctive invariant features (Lowe, 2004), histograms of oriented gradients (Dalal and Triggs, 2005), and SURF as a scale- and rotation-invariant detector and descriptor (Bay et al., 2006).

Radar-based techniques are also developed in the literature. Jahangir and Baker (2016, 2017) and Jahangir et al. (2017) developed a holographic radar using extended dwell Doppler characteristics to detect UAVs and distinguish them from birds. Drozdowicz et al. (2016) adopted an approach that uses multiple radars through the directional transmission to monitor a restricted area and used a geofence to detect UAVs. The mmWave radars detect UAVs at lower altitudes by using the Doppler spectrum. The mmWave radars detect moving UAVs concerning radars and prevent flight collisions of UAVs (Fioranelli et al., 2015; Nanzer and Chen, 2017; de Haag et al., 2016). However, mmWave radars cannot differentiate the UAV from objects with a similar shape, like birds. Thus, the UAV detection accuracy using mmWave radars is low.

Convolutional Neural Networks (CNN) are developed for the UAV detection. Saqib et al. (2017) study a transfer learning method using the combination of VGG16 and Faster-RCNN for the detection of UAVs. Al-Emadi et al. (2019) applied CNN, RNN, and RCNN architectures on audio-recorded samples to exploit the unique acoustic fingerprints of flying UAVs. Park et al. (2017) presented a method using Pan-Tilt-Zoom (PTZ) with a combination of camera and Faster R-CNN Inception ResNet algorithm for UAV detection with pan, tilt, and zoom actions. Choi and Oh (2018b) used Short-time Fourier Transform and Wigner-Ville distribution to transform micro-Doppler signatures from diverse UAVs to images. The optimal hyperparameters for the CNN were determined using a heuristic search. Several advanced trainable object detection models have been proposed to compete with these problems. The comprehensive literature has been conducted for the UAV detection and classification, which is given in Table 2.2.

Table 2.2: Comprehensive Literature Review on UAV Detection

Reference	Methods	Paper Type
Samaras et al. (2019)	Deep Learning based micro-Doppler (m-D) signature	Survey
Kang et al. (2020)	RF signal detection using Deep learning	Survey
Wagoner et al. (2017)	Image/video UAV detection using Deep learning	Survey
Taha and Shoufan (2019)	Video UAVs detection using Deep learning	Survey
Coluccia et al. (2020)	Deep Learning	Size, rotors, and payload-based UAV identification review
Stoica et al. (2020)	RSN Classifier	Noise signal-based UAV detection review
Rozantsev et al. (2017)	CNN and Boosted trees methods	UAV classification
Saqib et al. (2017)	VGG and ZF Fine-tuning	UAV classification
Dong and Zou (2017)	Feature Classifier	UAV classification
Peng et al. (2018)	Faster RCNN and ResNet 101 fine-tuning	UAV detection
Nalamati et al. (2019)	Faster R-CNN with ResNet-101	UAV detection
Lee et al. (2018)	Deep CNN	UAV classification
Hu et al. (2019)	Improved YOLO v3	UAV detection
Behera and Bazil Raj (2020)	YOLOv3	UAV detection
Li et al. (2021b)	UAV to UAV using 3-D CNN model	Single Stream
Seidaliyeva et al. (2020c)	ResNet-34 algorithm	Loaded and unloaded UAV images
Aker and Kalkan (2017)	YOLOv2	UAV detection
Seidaliyeva et al. (2020b)	YOLOv2	Loaded and unloaded UAV detection

(continued on next page)

Reference	Methods	Paper Type
Zhang and Chandramouli (2019)	RFCN and ResNet101	Detect and track UAVs
Schumann et al. (2017)	VGG-based network	UAVs and birds classification
Sommer et al. (2017)	Deep learning	UAV two-stage infrared target identification
Rozantsev et al. (2017)	CNNs	UAV detection with poor resolution
Magoulianitis et al. (2019)	Deep DCSCN model	Super-Resolution methods
Liu et al. (2021)	Deep Learning	UAV detection using single camera
Seidaliyeva et al. (2020a)	CNN	On-UAV video cameras
Behera and Bazil Raj (2020)	Deep learning algorithms	UAV detection via Background Removal
Lee et al. (2018)	U-Net network	Identify suspected UAV targets
Saqib et al. (2017)	Deep Learning	UAV detection using Depth maps
Craye and Ardjoune (2019)	Regression-based approach	Flying object detection using 3D spatio-temporal image cubes
Carrio et al. (2018)	Regression-based approach	Flying object detection using 3D spatio-temporal image cubes
Rozantsev et al. (2015)	Deep Learning	Markers-based UAV detection
Jin et al. (2019)	Deep Learning	Perception-based UAV detection
Lai and Huang (2020)	Improved YOLO	UAV monitoring and detection
Wu et al. (2018)	Deep Learning	Merged Doppler-based UAV detection

(continued on next page)

Reference	Methods	Paper Type
Kim et al. (2017a)	CNN	Super-resolution-based UAV identification and tracking
Li et al. (2022a)	YOLOv4	UAV and bird identification
Singha and Aydin (2021)	Deep Learning	Multi-camera UAV detection and tracking
Unlu et al. (2019)	CNN	Onboard markerless UAV detection and localization
Vrba et al. (2019)	CNN	Radiofrequency-based UAV detection
Basak et al. (2021)	Deep Residual Nets	UAV classification from RF fingerprints
Basak et al. (2022)	Deep Neural Network	RF signal-based UAV identification
Al-Emadi and Al-Senaid (2020)	DNN	Identify speed of takeoff and flying mode of UAV
Al-Sa'd et al. (2019)	CNN	Indoor UAV protocol detection
Sazdić-Jotić et al. (2022)	CNN + RNN	Indoor UAV protocol detection
Allahham et al. (2020)	DNN	Long RF signal audio-based UAV detection

Summary and limitations. Most prior works rely on a single modality (RGB or IR) or on sensing modalities (e.g., radar, audio) that either lack spatial detail or are brittle to clutter, according to Table 2.1 search databases and keywords. As a result, (i) day–night appearance shift, (ii) small apparent UAV size, and (iii) background similarity (birds, clouds, buildings) degrade recall and precision. This motivates a hybrid model that fuses complementary cues (e.g., appearance + motion/flow; RGB + IR) and explicitly targets robustness across illumination variations.

Therefore, single-stream approaches do not maintain high precision/recall across day–night conditions and diverse environments. Recent CNN-based detectors benefit from large-scale data and

architectures such as EfficientNet-B3 (Tan and Le, 2019). One-stage detectors (e.g., EfficientDet) couple such backbones with multi-scale feature fusion and focal loss to handle class imbalance. Building on these insights, we propose a hybrid model that trains on dual streams with data transformations (e.g., cosine-annealed schedules, focal loss) to improve robustness while keeping compute costs tractable for video.

2.2 Fusion Flow-enhanced Graph Pooling Residual Networks Background Study and Limitations

We next propose our Fusion Flow–enhanced Graph Neural approach relative to SotA detectors and resource-aware pipelines. For brevity, we emphasize methods that address small-object detection, motion cues, and challenging backgrounds. We refer to our family of models as Optical Flow-assisted Graph-Pooling Residual Network (OF-GPRN).

The literature encompasses several UAV detection methodologies based on RGB or IR images, leveraging Convolutional Neural Networks (CNN). Tian et al. (2023) introduced a YOLOv5-based detection paradigm tailored for small UAVs. This model refines detection by designing anchor box sizes, incorporating a convolutional block attention module (CBAM), and revising the loss function. Its prowess is accentuated by its capability to detect UAVs in complex and challenging environments. Similarly, Alsoliman et al. (2021) put forth a UAV detection approach that leverages random forest classification. This technique discerns patterns in video data to curtail the influx of packets emanating from UAVs. Later on, Alsoliman et al. (2023) propose a distinctive method introducing the notion of a pivot fingerprint model designed for pinpointing anchor packets within video streams for UAV detection. The framework of this model uses a two-tiered feature selection process, with the first phase being model-independent and the second phase being model-dependent.

Liu et al. (2020) studied an HR-YOLACT algorithm, which is an amalgamation of HRNet and YOLACT techniques. This model is architected to feature a lightweight prediction head, facilitating the detection of UAVs and extracting their features through instance-based semantic segmentation. Saqib et al. (2017), on the other hand, delved into a transfer learning technique for the identification of UAVs, using both VGG16 and Faster-RCNN. Meanwhile, Park et al. (2017) studied an approach that leverages a Pan-Tilt-Zoom camera system for UAV detection using the Faster R-CNN Inception ResNet algorithm.

Furthermore, Wei Xun et al. (2021) tailored the YOLOv3 model combined with transfer learning to detect UAVs using an RGB camera. This integration has further potential for real-time surveillance applications, especially on platforms like the NVIDIA Jetson TX2. In the same way, Reddy et al. (2023) showed how YOLOv3 could be used to find UAVs, highlighting how useful it is for effective monitoring in a variety of daytime situations. Lee et al. (2018) ventured into a machine

learning-centric approach, focusing on the identification of UAVs from RGB images. Their system keeps a vigilant eye on surveillance zones, pinpointing and cataloging UAVs using an RGB camera and subsequently determining their geographic position and manufacturer model.

Another approach was launched by Wang et al. (2022), consisting of a semi-supervised object detection technique termed Decoupled Teacher. Built on Faster-RCNN's foundation, this method employs unlabeled data with the aim of counteracting the imbalance between foreground and background observed in RGB camera feeds. Basak et al. (2022) described a YOLO-based UAV detection strategy that uses spectrogram images to classify and group spectral instances.

Choi and Oh (2018a) introduced various micro-Doppler signatures of UAVs. These were discerned using the short-time Fourier transform along with the Wigner-Ville distribution, both of which serve as tools to aid in the identification of UAVs. Multiple trainable object detection models were put to the test, comparing them with UAV identification tasks. Complementarily, Suh et al. (2023) researched UAV detection on platforms that are resource constrained, particularly focusing on video hardware. Their study meticulously melded algorithmic optimizations with FPGA hardware, aiming to adeptly scrutinize the intricacies of video streaming.

Qi et al. (2018) put forth a technique tailored to the recognition of consumer-grade UAVs, employing static infrared imagery. Their model uses an approach based on importance and integrates basic convolution, adaptive thresholding, linked domain filtering, and SVM-based discrimination. Sun et al. (2020) introduced TIB-Net, a specialized model designed for UAV detection through an RGB camera. This model, uniquely structured with a cyclic pathway, is particularly adept at detecting small-sized UAVs. Augmenting its efficacy, a spatial attention module is integrated, working to pare down data redundancy and extraneous noise. Further research, as presented by Diels et al. (2022) and Isaac-Medina et al. (2021a), showcases efforts wherein various CNN-based strategies are deployed to identify UAVs. These detection tasks are undertaken under a myriad of lighting scenarios, harnessing the power of RGB video feeds. However, despite the extensive study into CNN algorithms tailored for both RGB and IR-based UAV detection, their potency tends to be encumbered. This restriction results primarily from the difficult challenge of backgrounds with a high degree of similarity to the UAVs.

Summary and limitations. The prevailing models as depicted in the literature predominantly operate on single-stream data, either RGB or IR, during high background similarities, rather than processing fused optical flow images, where the visual presentation of UAVs and other entities can exhibit substantial variation between day and night vision scenarios. Additionally, the feature quality of the resultant frames is found to be considerably influenced by varying lighting conditions, which, in turn, often exacerbates the challenges due to the heightened similarities between the UAVs and their background environments. Moreover, prior detectors rely on convolutional backbones that operate on grid-structured pixels and can underutilize long-range relational cues when the target is small and camouflaged during higher background similarities. In our approach,

we practiced three steps to mitigate the problem faced by detection algorithms and improve the accuracy: (i) introduce fusion-flow features to emphasize motion-consistent regions across frames, (ii) build a graph over proposals/patches to capture relations (e.g., co-motion, context), and (iii) use residual graph pooling to preserve fine detail. This design directly addresses small-object saliency under background similarity.

2.3 Latency Efficient Edge Vision System Background Study and Limitations

Mobile Edge Computing (MEC), such as UAVs, is used for a wide spectrum of mobile vision applications, such as augmented reality (Hitchcock and Sung, 2018), trajectory planning (Li et al., 2023), or visual surveillance systems (to prevent problems in no-fly zones) (Noor et al., 2021; Ullah et al., 2023; Rezaee et al., 2022). However, the system should be accurate, real-time, and resource-efficient. To perform real-time tasks in computer vision, the UAVs should be capable of detecting large and small objects in dynamic environments with high accuracy. Moreover, UAVs are mostly used as MECs to stream videos to a server and get important data from the server after preprocessing to identify the objects (Galanopoulos et al., 2021; Xu et al., 2023). However, real-time video transmission and receiving tasks between UAVs and servers require a longer delay to transmit video frames and achieve high accuracy. As a result, neither their execution on resource-constrained user devices such as UAVs nor their transfer to remote cloud servers is practicable because of higher transmission delay and low accuracy (Dong et al., 2022). Moreover, monitoring no-fly zone areas and capturing suspicious objects, a huge volume of live video transmission by UAV required as an edge device to the server will increase the consumption of computing resources (Han et al., 2024), (Khan et al., 2022). Researchers introduced different system models to optimize the video transmission latency.

Dong et al. (2022) introduced an ultra-low latency frame delivery video streaming model that uses the UDP-based QUIC protocol and the widely compatible TS format. The model is constrained with a lower latency. Wang et al. (2021a) presented an EdgeDuet model that divides the frames into tiles and offloads small object detection to the edge in high resolutions for the detection and tracking. While Zhang et al. (2022a) worked on the edge instance segmentation technique (EdgeIS) to segment the region of interest on the edge and transfer it to the mobile device in low latency. The EdgeIS model detects objects and tracks their motions with mask transferring. Moreover, EdgeIs models utilize the contour-instructed edge inference acceleration scheme to reduce the latency. Yang et al. (2022) presented the Flexpatch model that utilizes an object tracker and patch recommender along with optical flow for the detection of small objects with a minimum delay and to address occlusion, changes in object appearance, and the presence of additional objects. Wang et al. (2019) studied feature-based video transmission, using Lagrangian dual decomposition to optimize the transmission latency and semantic image segmentation for video feature selection. Ghosh et al. (2023) studied edge and cloud collaborative work to use heavy deep

learning models in the cloud and small models at the edge to fuse their predictions to optimize latency. Liu and Jiang (2024) designed a multi-approach model by extended splitting-merging streaming for the optimal latency and multi-link retransmission to reduce the packet loss. Yaqoob et al. (2024) presented a DRL-based soft actor-critic model to improve the quality of experience of video transmission latency from UAV to the ground server, while Qu et al. (2024) studied UAV-Swarm-Aided modeling based on a disaster response platform for the video transmission from UAVs to the ground as air-to-ground coordination. Similarly, Khan et al. (2024) presented a public safety communication network model that uses observation UAVs to receive the video streams from affected area ground users and transmit them to the nearby functional ground base station. Song et al. (2024) introduced an AI-driven, multipath transmission using a UAV live streaming system to optimize bandwidth aggregation and transmission reliability. In the meantime, Wu et al. (2024a) introduced a multiobjective control strategy learning for the UAV using a DRL algorithm called soft actor-critic to manage the unstable transmission quality problem of videos in a dynamic topology network.

A region-of-interest-based (RoI) sub-area of video frame transmission appears to be a solution for meeting low latency and high accuracy criteria, as it allows users to send their video frames to nearby servers equipped with GPUs for processing by deep learning models (Wang et al., 2021a; Zhang et al., 2022a; Yang et al., 2022). In this case, various sub-areas of video frames can be offloaded from a UAV device using existing edge computing systems. However, user-side parameters like image resolution and encoding rate and server-side parameters like NN architecture have a big effect on video analysis if we do not consider how the pixels of the RoI correlate with the pixels of the video frame background environment (Zhang et al., 2022a). Wang et al. (2021a) studied a split-based object detection model (EdgeDuet) between UAV devices and the edge. They offloaded the small objects onto the server. The RoI of the frame is used to optimize small objects and content-prioritized tiles for offloading. The RoI of the frame transmitted with pixel blocks potentially contains small objects of high resolution. The rest of the frame has large objects transferred to low resolutions and processed on UAV devices. However, EdueDuet is based on tile selection, which makes tiles have an unnecessary area that is not related to small objects for the transmission and increases the transmission latency. Moreover, there might be a server detection accuracy problem due to the tile approach, as shown in Fig. 2.1. Yang et al. (2022) introduced a Flexpatch model that allows for the detection of small objects with a small delay while also addressing occlusion, changes in object appearance, and the presence of additional objects. However, the Flexpatch model works with static cameras due to its optical flow-based system modeling and it is unsuited for dynamic-based vision systems like UAVs.

Summary and limitations. Existing systems reduce end-to-end delay via protocol tweaks, tiling or ROI selection, and edge–cloud collaboration, but they often (i) forward redundant background pixels, (ii) ignore task-aware importance (small-target AP), or (iii) degrade under bandwidth or packet-loss bursts. Our objective is to co-optimize what to transmit (task-aware ROIs or features) and how to transmit (adaptive scheduling/aggregation), targeting both latency and constraint-based

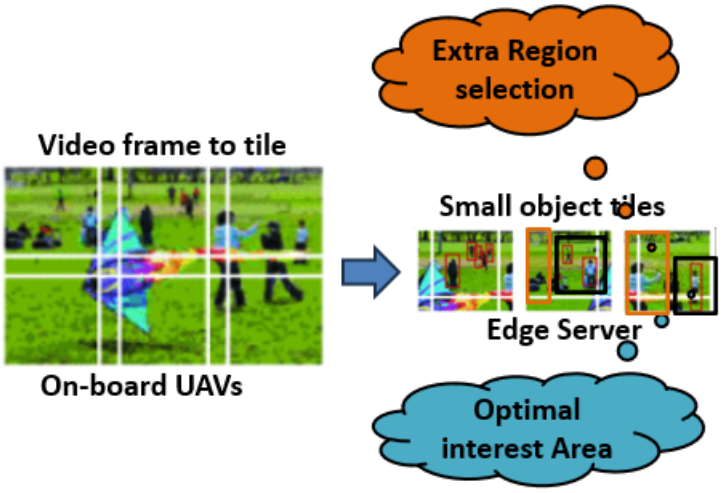


Figure 2.1: EdgeDuet non-optimal region selection for the transmission of small objects to server.

average precision for UAV detection.

2.4 Technical Pillars

In this section we provide a brief introduction to the emerging AI developments referred to above, which are key pillars for our thesis. In Section 2.4.1 we discuss the reasons why the feature pyramid network of hybrid deep learning is significant for object detection. In Section 2.4.2, we present the graph convolutional networks that are used for image data. In Section 2.4.3, we discuss reinforcement learning as it is applicable to the optimization of latency.

2.4.1 Feature Pyramid Network

Training hybrid deep learning models with multimodal inputs such as RGB and infrared data may introduce cross-modal inconsistencies that complicate feature fusion. To effectively handle objects at different scales, the Feature Pyramid Network (FPN) utilizes the hierarchical feature maps of convolutional networks to construct a multi-scale feature representation with strong semantic information at all pyramid levels (Lin et al., 2017). Mostly FPNs are used for region proposal and detection tasks (Zhao et al., 2019, 2021; Wang and Zhong, 2021; Zhu et al., 2022; Zhang et al., 2022b; Deng et al., 2022; Yang et al., 2023). In our context, FPN-style multi-scale fusion is important because UAVs appear at variable sizes and often occupy only a small number of pixels. By sharing detection heads across pyramid levels and aggregating features from low and high resolutions, the detector becomes more robust to scale and illumination changes in day–night settings.

FPNs come up with the idea to use a bottom-up pathway, a top-down pathway, and lateral connections to combine features of different resolutions and modalities. The method makes multi-scale

feature maps from single-scale input images, and it doesn't depend on a specific backbone architecture. The model focuses on simple standard convolutional layers and shares classifiers at all levels of the pyramid network. However, the model structure depends on the complexity of the objective and might be used with more complex and deep convolutional layers. Modern extensions such as BiFPN introduce learnable fusion weights and repeated bidirectional connections, which we leverage when building the hybrid model. We also consider training strategies (e.g., cosine annealing and photometric and geometric augmentations) to encourage stable convergence when mixing RGB and IR data.

Researchers have proposed several models and extensions to the standard FPN architecture to address its limitations and enhance feature representation. There are several improved models studied to address the limitation of the standard FPN model for object detection. Among these, Liu et al. (2018) introduced a path-based aggregation network to add augmentation from a bottom-up path and improve the information flow with localization capability for FPNs. Tan et al. (2020) studied the Bi-directional Feature Pyramid Network (BiFPN) for the EfficientDet model to allow learnable weights in easy and fast multi-scale feature fusion to detect the objects. Zhong et al. (2022) introduced the image pyramid strategy for anomaly detection instead of conventional patch-level error-based anomaly evaluation. The model predicts error maps at multiple scales and aggregates the maximum errors across all scales while avoiding the need for selecting an optimal patch size. Wang et al. (2020) studied attention mechanisms and integrated them with FPNs that adaptively adjust channel-wise features. All these models support active research in FPNs to improve object detection effectiveness and overcome limitations. In Chapter 3, we discuss the configuration choices that work best for our dual-vision setting and provide ablations to quantify their effect on small-object accuracy.

Zhao et al. (2021) presented a graph feature pyramid network update to the feature learning and information at multi-scale interactions from fixed to intrinsic and hierarchical structure features during object detection. Deng et al. (2022) used a feature texture transfer module to extend the feature pyramid network for regional textures and deep semantics. The model employs a cross-resolution distillation strategy and a foreground-background-balanced strategy to learn small object features from large input. Yang et al. (2023) introduced the asymptotic feature pyramid network (AFPN), which is used to avoid large semantic gaps in features during extraction and preserve information by gradually fusing low- and high-level features. AFPN filters features using adaptive spatial fusion with multi-level feature fusion to improve detection accuracy. Feature learning using a single, top-down pathway is limited by simple merging strategies and single-track information flow, and using bidirectional parallel feature pyramid networks, which utilize cross-level channel attention refinement and parallel feature aggregation blocks, makes it possible to enable richer multi-scale feature extraction from each level Zhang et al. (2022b). These advances inform our design of the hybrid model and the graph-enhanced network in later chapters.

These developments of FPN models are important for robust multi-modal feature learning, adaptability to new data, object detection, and efficiency. However, there are several limitations of

standard FPNs, and the integration of new approaches such as hybrid models, attention modules, adaptive fusion, and hierarchical structures can make them capable and efficient with complex, multi-scale, and diverse data feature learning models. The innovations in FPNs can make it applicable for new domains in computer vision to achieve state-of-the-art results. We therefore study an FPN/BiFPN-style backbone as a practical foundation for our hybrid detector and use it again inside our graph-based refinement network.

2.4.2 Graph Neural Networks

The contextual correlation and relationships between detection object pixels, patches, and object proposals are still a challenge for object detection models. The detection model's spatial and semantic connection lacks structured reasoning. Graph Neural Convolutional Networks (GCNs) are studied to integrate into detection models and use graph nodes for regions and objects with adjacency to encode the relation between the pixels and regions to enhance contextual reasoning. In aerial scenarios, where UAVs are small and sometimes invisible, these relationships (co-motion, proximity, and background stability) provide significant indicators. By constructing graphs based on superpixels or proposals and enabling interaction between nodes, the model might reduce false positives while emphasizing consistent motion and appearance.

Li et al. (2020) studied graph-assisted reasoning with the help of graph nodes and edges to reduce the false positive rate of object detection proposal scores when using standard convolutional neural network models, while Zhao et al. (2021) introduced a graph feature pyramid network to propagate the contextual information across scales. Zhang et al. (2023b) studied graph fusion networks for the aerial images and focused on rotating the bounding boxes. The model is applied to suppress redundant detections, which makes clustering of proposals in subgraphs. Han et al. (2022) studied to go even further and introduced Vision GNN using patch-wise message passing instead of convolutional neural network layers, while Kim et al. (2022) presented an uncertainty-aware graph network for aerial images that utilizes semantic-spatial graph reasoning for object detection. Building on these ideas, our OF-GPRN uses residual pooling to avoid losing fine details when aggregating information on the graph, and it uses optical flow to define motion-aware edges that better capture true object relations in videos.

Similarly, there are several studies that use graph neural networks for human-object interaction detection. Zhang et al. (2023a) introduced interactivity recognition-based graph neural networks, which are bipartite graphs between humans and objects. The model adopts attention-enhanced message passing and modeling relationships of pose and proximity to detect human and object interactions, while Ulutan et al. (2020) presented a visual semantic graph for human-object interaction that uses graph reasoning with the help of anchored person-object pair analyses. Wang et al. (2021b) presented graph neural networks for encoding temporal consistency between frames of videos and multi-object detection. Additionally, Weng et al. (2020) studied graph neural networks to fuse two different modality data appearances and motions. Shi and Rajkumar (2020) introduced point cloud data to detect object boxes utilizing graph neural networks, which create graphs from

point cloud data, while sparse voxel-graph attention network is introduced by He et al. (2022) to enhance feature learning with graph-based local and global attention mechanisms. Xiong et al. (2023) studied density clustering with graph neural networks using multi-scale hierarchical design and feature fusion for point set partition, which captured local and global features. Despite these works addressing different research topics, the basic concept of reasoning over structured relations is immediately useful for UAV surveillance in complex environments.

These graph neural network models might improve detection accuracy using multi-hop feature propagation, adaptive context fusion, and structural reasoning. However, modeling complex relationships of object pixels in complex environmental effects creates difficulties for scalable graph construction and facilitates better aggregation and across-node information aggregations. Using enhanced scalability and adaptive mechanisms can allow GNN to address diverse applications in computer vision for complex environmental effects. We therefore design the Graph Residual Split Attention Network (GRSaN) to be precise and residual, so it can run within complex environmental effects and capture useful context.

2.4.3 Reinforcement Learning

RL made it easier to minimize on-board latency for mobile edge devices and improve the quality of dynamic video communication and computations. Recently, reinforcement learning has been used for different on-board mobile edge devices to optimize the transmission latency, buffering, and inference for video and image UAV systems (Uitto and Heikkinen, 2021; Masaracchia et al., 2021; Sharma et al., 2023, 2024; Gupta et al., 2016). In our problem, the key decision is *what* to transmit (full frames versus task-aware regions) and *when/how* to transmit under variable bandwidth, packet loss, and compute availability. RL is a natural fit because it learns policies that trade off accuracy with latency in real, changing environments.

Xiao et al. (2020) introduced reinforcement learning with an adaptive bitrate model, which collects on-board data with the help of quantization preprocessing. Moreover, the model uses LSTM for handling noise and preserving temporal throughput features through observation. Burhanuddin et al. (2022) studied a dynamic communication model that uses Actor-Critic and DQN for UAV video streaming. The model adapts to fluctuating channel conditions and addresses the quality of experience to optimize the UAV position, video resolutions, and video transmission for dynamic wildfire monitoring. Yun et al. (2021) presented centralized training with decentralized execution of a multi-agent reinforcement learning model for low-latency UAV communication. The model is used to develop symmetric and inter-agent communication of UAV data and coordinated transmission without collisions or delays. Dhuheir et al. (2023) used PPO to distribute the convolutional neural network layers and optimize the end-to-end video latency for multi-UAV surveillance. The model analyzes the UAV's mobility and resources and defines the trajectory planning in UAV swarms for video transmission. Xiao et al. (2021) studied model-free reinforcement learning to improve the quality of experience for UAV anti-jamming video transmission. The model optimizes the quantization parameter along with the channel coding and transmission outages to reduce the

energy consumption without past knowledge. Souissi et al. (2024) introduced end-to-end deep reinforcement learning to track UAV to UAV using an RGB camera under dynamic, high-speed, and low-light conditions. The model maps raw events to flight-control actions for target UAV tracking and minimizes latency. We adopt an A2C backbone and incorporate a GCN that groups pixel-correlated regions. The agent selects the region(s) to transmit such that end-to-end latency is minimized while the detection accuracy remains above a configurable threshold.

Wu et al. (2021) presented a deep deterministic policy gradient model designed to optimize sensing networks by coordinating UAV trajectories and scheduling decisions for transmitting data from UAVs to base stations or mobile devices. The model used the Markov Decision Process formulation technique for the problem. Wu et al. (2024a) presented the Soft Actor-Critic model to maximize the quality of experience and minimize the on-board energy consumption under dynamic network conditions. The model optimizes the UAV trajectory and selects different video layers and allocates the bandwidth during transmission. Miao et al. (2024) studied the deep deterministic policy gradient algorithm to optimize the power allocation and minimize the transmission latency by applying the Markov decision process problem formulation for UAV trajectory. Moreover, to prevent the UAV eavesdroppers, which pose a security threat when transmitting the video to the MEC system due to unpredictable positioning, Zhao et al. (2024a) presented a deep reinforcement learning model to transmit video to the MEC system and optimize the resource allocation, video resolutions, and UAV positions. Li et al. (2025c) used random distribution of user devices to optimize data offloading latency and trajectory planning with a deep reinforcement learning approach, while Xu et al. (2025) presented a constrained deep reinforcement learning algorithm to optimize long-term inference delay with neural network model selection for UAV and IoT communication under jointly optimized task assignment. Integrated video distortion with physical-layer challenges might be jointly optimized by video resolution encoding, computational offloading, and duality of experience with security using a hierarchical reward-based deep reinforcement learning model (Zhao et al., 2022). In Chapter 5, we formalize the objective and report latency/accuracy trade-offs under bandwidth and loss variations, comparing with recent edge vision baselines.

2.5 Chapter Summary and Concluding Remarks

This chapter reviewed (1) single-stream RGB/IR and classical sensing approaches for UAV detection, (2) SotA deep detectors and resource-aware pipelines, and (3) latency-optimized streaming for edge vision. Across these lines of work, three persistent gaps emerge: (G1) robustness across day–night conditions and background similarity; (G2) reliable small-object detection leveraging motion/relational context; and (G3) joint optimization of transmission latency with task-level accuracy. The next chapters address these gaps: Chapter 3 introduces our hybrid deep learning model for dual-stream training; Chapter 4 details the Fusion Flow–enhanced Graph Pooling Residual Networks; and Chapter 5 presents a graph neural network–assisted DRL framework for video transmission latency optimization.

Chapter 3

A Hybrid Deep Learning Model for UAVs Detection in Day and Night Dual Visions

In this chapter we explore **Challenge 1-** hybrid model for day-night visions as already discussed in Section 1.2.1. We developed a hybrid deep learning model for processing RGB and IR videos, which is a non-trivial since the RGB and IR video frames are trained independently with day or night vision features, as shown in Figure 3.1. As a result, RGB and infrared video streams cannot always be directly integrated for UAV detection in a dual-vision framework because the two modalities exhibit different visual characteristics, including illumination sensitivity and spatial alignment. Consequently, developing a hybrid model may increase system complexity, as cross-modal feature inconsistencies can arise during the joint training of RGB and IR data.

In the hybrid deep learning model, the UAV needs to be located in the RGB and IR videos and differentiated from other objects. The position of the UAV can be located by using bounding box localization (BBL), which determines the height, width, and X and Y-coordinates in annotation form and avoids the annotation errors (He et al., 2019). Specifically, correct annotation is typically used to determine the UAV position with annotation of the upper left corner and the lower right corner of the UAV. Since the background color can be similar to the color of the UAV, the annotation with BBL experiences annotation errors, which results in classification and regression losses in the deep learning of BBL. In particular, the classification loss is due to the color difference between the UAV and the background, while the regression loss defines the difference between the actual bounding box and the predicted one of the training. In addition, the classification and regression loss of the BBL training depends on the position of the UAV in the video and the annotation error.

Therefore we propose a new hybrid deep learning model for the UAV detection with day and night dual visions. The contributions of our proposed hybrid deep learning model work are as follows:

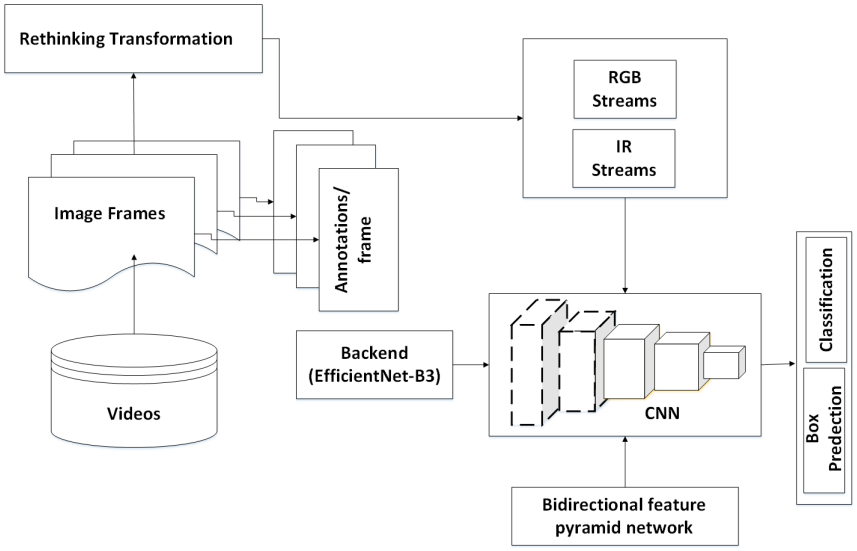


Figure 3.1: The proposed architecture for the UAV detection using dual streams.

- We propose a new hybrid deep learning model to improve the detection accuracy of the UAV in the day and night dual visions. The hybrid deep learning model adopts photometric and geometric augmentation to accelerate the training model's convergence and reduce the classification and regression losses. Moreover, the proposed hybrid model develops cosine annealing, which freezes stepwise the initial layers of the deep learning model to reduce the training time.
- The proposed hybrid model is developed to enhance the training of the BBL by using the information of coordinate dimensions (height and width) to determine the actual location of the UAV and reduce the annotation errors of the UAV detection.
- To evaluate the proposed hybrid model in the real-world, experiments are conducted to detect the UAV in daytime and nighttime. The experimental results show that the proposed hybrid deep learning model achieves 65% higher detection accuracy than the benchmark EfficientDet. In addition, the convergence of the hybrid model is 10% faster than the EfficientDet.

The remaining sections of the chapter are organized as follows. In Section 3.1, we investigate the proposed hybrid deep learning model and its implementation. We present the experimental setup and performance evaluation in Section 3.2. Finally, we summarized and concluded with remarks on the chapter in Section 3.3.

3.1 Hybrid Deep Learning System Model

3.1.1 Dataset Preparation and Preprocessing

In this chapter, we utilize open-source Anti-UAV catch video datasets (Jiang et al., 2021). This dataset contains 160 HD dual streams (RGB and IR videos), each with 100 validations and 60 training videos. Each video has multiple UAV models with three different sizes (large, medium, and tiny). The video frames contain different backgrounds: day/night, cloudy/clear, similar objects (buildings, birds), scale variations, and occlusion, as shown in Figure 3.2. The UAVs move at different speeds and abruptly stop at very high speeds (100 mph). We used the 100 videos from each stream for the training of our model. 90 of the videos are used for training, and 10 videos are for validation/testing of the proposed model.

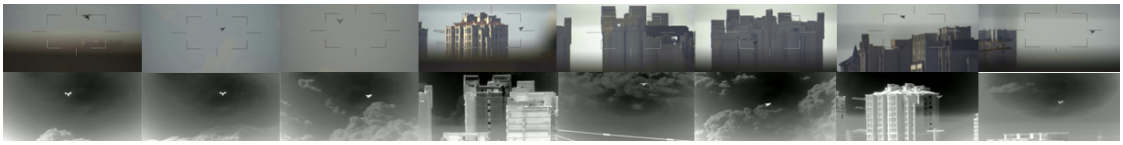


Figure 3.2: The first row is of RGB video frames, and the second row is of IR video frames from different videos and background variations of different sizes of UAVs.

Annotation errors in video data for computer vision object tracking tasks degrade the model performance in the case of shifted, missing, or inaccurate bounding boxes by human labeling. The trained model can misclassify UAVs during testing or deployment. There is template matching between the consecutive video frames bounding box displacement. It is essential to remove the linear drift and displacements to refine the object annotations before training the system model. Although the video frames of the Anti-UAV catch video datasets (Jiang et al., 2021) are labeled, the labeling undergoes high annotation errors, as shown in Figure 3.3 due to the poor UAV detection performance (Koksal et al., 2020). To ensure the higher accuracy of the system model, we



Figure 3.3: Green bounding boxes, such as additional boxes, missing boxes, and shifted boxes, indicate anti-UAV dataset annotation errors. While red bounding boxes show the correct version.

need to have an algorithm that correct the annotations and compares the original and corrected annotations by computing drift corrected displacements of bounding boxes.

3.1.1.1 UAV Localization

The dataset annotation needs the proper conversion by using the information of coordinate dimensions. During the flight process, the location of the UAV can be determined by the X-coordinates

and Y-coordinates with height (H) and width (W) of the UAV to reduce annotation errors in the detection. Therefore, we detect four ($x_{\max}$, $x_{\min}$ and $y_{\max}$, $y_{\min}$) translation parameters, which detect the UAV with the top-left corner and bottom-right corner. The X-axis position of the UAV presents the width ($x_{\max}$-$x_{\min}$) of the UAV, while Y-axis represents the height ($y_{\max}$-$y_{\min}$) of the UAV. The UAV's position varies by altering its rotation by global coordinate ($x_{\max}$, $x_{\min}$ and $y_{\max}$, $y_{\min}$ pivoted) and is not fixed horizontally (non-pivoted), so the width and height change according to the sum of the X-axis ($x_{\max}$+$x_{\min}$) and Y-axis ($y_{\max}$+$y_{\min}$) respectively. In our work, the UAV is correctly located in RGB video frames and thermal (IR) video frames. To decide whether the operating procedure is in RGB or IR, we used the brightness information provided by the surveillance camera. The proposed UAV localization method tracks the UAV's movement multiple times for accurate localization, as illustrated in Algorithm 1.

Algorithm 1 is the improved model of Koksal et al. (2020), which set a search range for template matching with a brightness threshold of greater than 100 for RGB. Thereafter, cross-correlation is used with the video frame k+1 estimated UAV position from the center of the annotated k video frame. The annotation errors of k+1 and k are defined by computing displacement $\vec{u}_{k+1}$. Subtract the cumulative displacements and use the fitted linear trend to calculate the drift-corrected displacement between k+1 and k. Lastly, we checked and compared the refined annotations with the original and corrected annotations to add or reject. Table 3.1 presents the corrected dataset that reflects the improvements made in UAV localization.

Table 3.1: UAV dataset statistical approach of the RGB and IR video frames.

Categories	Training set	Validation set	Testing set	Total
Number of video frames	162k	19k	19k	200k
Percentage (%)	81.0%	9.5%	9.5%	100%
Number of RGB video frames	81k	9.5k	9.5k	100k
Number of IR video frames	81k	9.5k	9.5k	100k

3.1.2 Backbone of the UAV detection

Pixels-to-features translation is developed with the proposed hybrid deep learning model. For scaling detection of the UAV, our hybrid model automatically adjusts the scaling of the width, depth, and resolution of the CNN consistently to provide different sizes of the UAV, as shown in Figure 3.4. Moreover, the hybrid model extends EfficientNet to enable features extractor for the UAV detection to achieve high accuracy and inference speed (Tan and Le, 2019). EfficientNet effectively scaled layers using multi-objective neural architecture search, which optimizes accuracy and computational cost in floating-point operations per second (FLOPS) in $\mathrm{ACC}(m) \times \left(\frac{\mathrm{FLOPS}(m)}{T}\right)^{w}$. The accuracy ACC and FLOPS divided by the target 400M FLOPS with the scaling trade-off controlling the hyperparameter $w = -0.07$ of model m is used as the optimization goal. The model has a special feature building block of mobile inverted bottleneck convolution (MBConv) and squeeze-and-excitation for optimizing the accuracy. Moreover, EfficientNet assumes a fixed

Algorithm 1 Template Matching for Correction of Temporal Annotation and UAV Localization

Require: video_frames, init_annotations, brightness
Ensure: refined_annotations
Set *search_range* for template matching
Initialize *refined_annotations* $\leftarrow$ empty list
for each video in video_frames **do**
    if brightness $>$ threshold **then**
        video_type $\leftarrow$ RGB
    else
        video_type $\leftarrow$ IR
    end if
    Initialize corrected_annotations $\leftarrow$ empty list
    Initialize cumulative_displacement $\leftarrow (0,0)$
    for each video_frame index k from 0 to $N-2$ **do**
        Extract $(x_{\min}, x_{\max}, y_{\min}, y_{\max})$ from annotation
        $W \leftarrow x_{\max} - x_{\min}$
        $H \leftarrow y_{\max} - y_{\min}$
        $(x_c, y_c) \leftarrow \left(\frac{x_{\min}+x_{\max}}{2}, \frac{y_{\min}+y_{\max}}{2}\right)$
        Adjust for rotation/pivoting using $(x_{\min}+x_{\max}, y_{\min}+y_{\max})$
        Extract template from video_frame k using bounding box
        Using *search_range* to perform template matching in video_frame $k+1$ around (x_c, y_c)
        Compute displacement $\vec{u}_{k+1}$ from match
        cumulative_displacement $\leftarrow$ cumulative_displacement $+ \vec{u}_{k+1}$
        Store cumulative_displacement_list$[k+1] \leftarrow$ cumulative_displacement
    end for
    To cumulative_displacement_list fit linear trend
    Compute drift_corrected_displacements $\leftarrow$ cumulative_displacement_list $-$ trend line
    for each video_frame index k **do**
        Adjust object center using drift_corrected displacement
        Recalculate bounding box from new center and (W,H)
        Append corrected annotation to corrected_annotations
    end for
    Compare original vs corrected annotations
    if human operator prefers corrected **then**
        Add corrected_annotations to refined_annotations
    else
        Add original_annotations to refined_annotations
    end if
end for
return refined_annotations

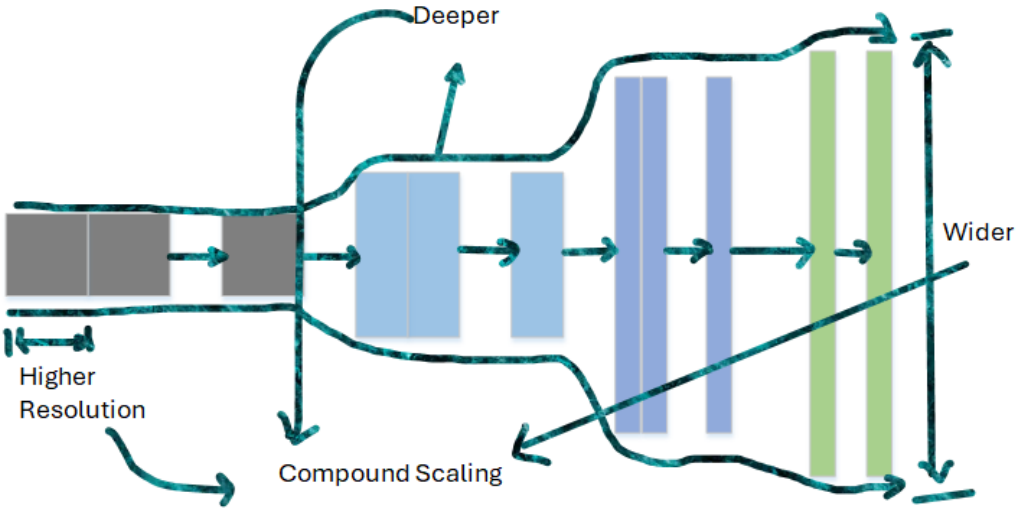


Figure 3.4: The compound scaling of width, depth, and resolution uniformly in a fixed ratio.

scaling coefficient with $\phi = 3$ for EfficientNetB3 with depth scale of $\alpha = 1.2^{\phi}$, width scale of $\beta = 1.1^{\phi}$ and a resolution scaling factor of $\gamma = 1.15^{\phi} \cdot 224$ under the constraint of $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$. To obtain different models of EfficientNet from B1 to B7 depends on the ϕ and that avoids high computational cost.

3.1.3 Dual-stream video processing

Using a hybrid deep learning model system to process dual-stream videos is a challenging problem for UAV detection. Most UAV detection techniques are based on single-stream video, i.e., either RGB or IR. Moreover, cross-modality uses dual-stream deep learning structures to process both RGB and IR inputs. During multimodal training, RGB features may dominate infrared features due to differences in modality characteristics and information representation. To mitigate this imbalance, we design a hybrid deep learning model based on the EfficientDet architecture, which enables effective multi-scale feature fusion and balanced feature learning across the dual streams. As a result, the model achieves more stable optimization and improved detection performance during training (Tan et al., 2020). EfficientDet is used for object detection to predict the object class and draw a bounding box around it. The model's performance is backed by ImageNet-pretrained EfficientNets, which balance the network depth, width, and resolutions in a systematic way of scaling that leads to high accuracy even with both RGB and IR data training in a single phase. However, the EfficientNet is not only capable of leading the model to better performance, but EfficientDet also studied top-down and bottom-up bidirectional feature fusion as Bidirectional Feature Pyramid Network (BiFPN), as shown in Figure 3.5.

In the proposed hybrid deep learning model system, we investigate BiFPN, which is an extension to EfficientDet. Specifically, BiFPN combines the ideas from Feature Pyramid Network, Path Aggregation Network, and NAS-FPN in the form of multi-level feature fusion, which efficiently helps the proposed hybrid model extract features for both RGB and IR streaming. The proposed

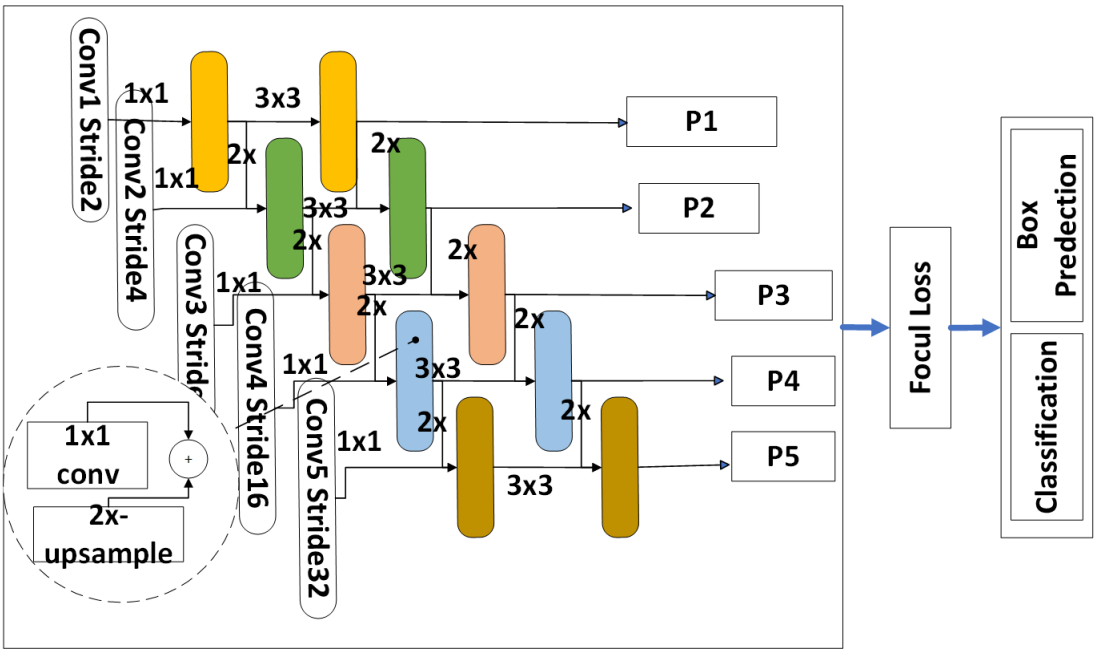


Figure 3.5: A top-down and bottom-up bidirectional approach to feature fusion, ensuring accurate object classification and bounding box prediction.

hybrid model enables efficient processing of dual-stream inputs while reducing computational complexity and supporting stable feature learning during training. The proposed hybrid model takes advantage of the BiFPN to minimize the computational cost by removing nodes with a single input edge and adding extra borders from input to output. Moreover, the hybrid model utilizes BiFPN to fuse high-level features of RGB and IR efficiently. The BiFPN configures additional weight to the input features, enabling the hybrid model to learn the feature importance of each input.

3.1.4 Photometric and Geometric Augmentation

The photometric and geometric augmentation shown in Figure 3.6 aims to enhance the accuracy and efficiency of the proposed hybrid model (Zhang and Ma, 2022). Since the CNN architectures are prone to color, rotation, and an axis-aligned bounding box transformation, we study the photometric and geometric augmentation in color background transformation, i.e., contrast, flapping, equalizing, solarizing, and sharpness, to apply the visual effect to video frames. In the photometric and geometric augmentation, the random probability for all parameters is set to 0.5, except 0.1 for solarization with a threshold of 128. Solarization tonal the values of the video frames in which dark areas appear bright and vice versa. The factor for adjusting the color is $0 \leq x \leq 2$ to maintain the originality and quality of the data. The photometric and geometric augmentation rotation of the UAV performed 45-degree clockwise and anti-clockwise can vary according to the rotation degree from the center (width and height by 2) of the UAV. The new bounding box dimensions of the video frames have been computed after getting sine and cosine from the photometric and geomet-

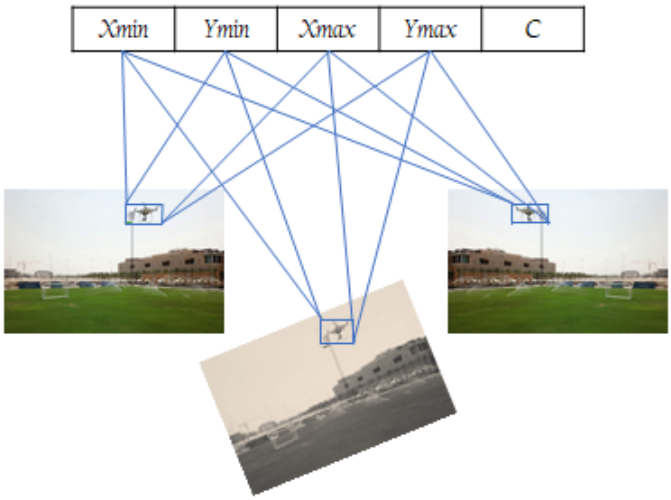


Figure 3.6: Photometric and geometric augmentation of UAV with a boundary box for precise bounding box localization.

ric augmentation rotation matrix. The bounding box annotation of the photometric and geometric augmentation is tackled by an array of $O \times 5$, where O is an object (i.e., the UAV) in the video frame, and 5 gives the attributes. The attributes include top-left corner coordinates, bottom-right corner coordinates, and a class of UAVs. In addition, the new boundary box of the rotation has been translated to the integer scaling using the photometric and geometric augmentation to avoid the missing color problems. The missing color is due to rotation in which new pixels are added to the boundary of the video frames, and the new ones replace the original pixels. The photometric and geometric augmentation determines the nearest neighbor (integer scaling) is an up-scaling method of the coordinate mapping to the nearest pixel value.

3.1.5 Cosine Annealing

The proposed hybrid deep learning model reduces the training time using cosine annealing. Initial layers of the CNN represent low-level features. These early layers take up most of the allocation but have updates of very few parameters and converge to straightforward configurations influences compared to later layers (high-level feature layers), where most of the parameters are updated during training. The initial layers do not require as much fine-tuning as influenced by this consideration. Cosine annealing is developed with the hybrid model to freeze the layers' weights, which preserves the pre-trained features for accelerating the learning. Cosine annealing is used for freezing out stochastic depth (Brock et al., 2017; Huang et al., 2016). In particular, cosine annealing trains each layer for several runs, gradually "freezes out" layers using equation 3.1, and excludes them from backward passes to accelerate the hybrid model.

$$LearningRate(\alpha_i) = 0.5 * \frac{\alpha}{t_i}(1 + cos(\frac{\pi * t}{t_i})) \tag{3.1}$$

where α defines the learning rate with the initial learning rate zero. t_i is the number of iterations between a user-selected t_0 and the total iteration during training. Iteration (t) depends on the rule of linear scheduling to reduce the training time.

3.1.6 Focal Loss

The two-stage object detection is computationally expensive to find the candidate object box (Lin et al., 2017; Hou et al., 2019; Liu et al., 2018). First, the two-stage object detectors use the Region Proposal Network to predict candidate bounding boxes. In the second stage, features are pooled from each candidate box for classification and bounding box regression tasks using RoI (Region of Interest) operations. One-stage object detectors (i.e., EfficientDet) can detect the UAV without creating proposal drawings (region proposals) (Tan et al., 2020). However, the problems of instances on new data mainly exist in the one-stage neural network detectors. The prediction of the deep learning model is far from the correct prediction if not used the focal loss. Focal loss is specially designed for the one-stage detection scenario. It can tackle the significant imbalance between foregrounding and background UAVs for the proposed training hybrid model to reduce the weight toward correct prediction. Cross-entropy is a frequently adopted loss function due to its high precision in compare the approximate models.

$$Cross\ Entropy(_{p,t}) = \begin{cases} -\alpha_t\ log(p), \text{ if } t = 1 \\ -\alpha_t\ log(1-p), \text{ otherwise} \\ \text{When, } t \in \{\pm 1\}, p \in [0,1] \end{cases} \tag{3.2}$$

In 4.7, the t is the target value as ground truth, and the p is the probabilistic estimated model value for the UAV detection. Where α_t is the balanced parameters for positive and negative examples, but it can not distinguish between easy and hard examples. The modulated factor $(1-p_t)^{\gamma}$ of the focal loss is necessary for the numerical stability by using the downweight methodology. The 4.7 is updated in the focal loss as: beginequation

$$Focal\ Loss(_{p,t}) = -\alpha_t\ (1-p_t)^{\gamma}\ log(p_t) \tag{3.3}$$

The focal loss down-weighted tuning process is dependent on the γ, and it varies from 0 to 2 to adjust the rate of easy examples. If γ is 0, then the focal loss equals the cross-entropy. The UAV detection scenarios increase the γ up to 2 to obtain the best training result.

3.2 Experiments and Results

The experiments of the UAV detection are performed on the UAV dual-stream dataset of the Anti-UAV CVPR workshop. The dataset is divided into 162k video frames for training and 19k video frames for testing. Each video frame is labeled with boundary boxes for UAVs and correctly localized by the UAV localization technique. GPU used a workstation equipped with a GeForce RTX

2070 to train the model using Keras with the backend of Tensorflow. We opted for EfficientDet compatible with training on the GeForce RTX 2070 with a batch size of 8. The Adam optimizer is used for its fast and easy convergence to the optimal point concisely compared to other optimizers. The learning rate decay is set to 0.001 with $\beta 1$ of 0.9 and $\beta 2$ equal to 0.999. The model trained for 100 epochs, which took 97 hours. The focal loss kept the same as in the original EfficientDet model α=0.25 and γ=1.5 with an aspect ratio between 1/2 and 2.

3.2.1 Results

We checked our hybrid deep learning model on the ground and with test videos to emphasize efficiency in various UAV positions. First of all, let us evaluate the identification accuracy of the proposed hybrid model with many traditional assessment methods. The video used in the experimentation is not used for testing or validation. We have also attempted to test the hybrid model for all different backgrounds and UAVs. During experimental results, different parameters are used to evaluate the performance of the training hybrid model. Focal loss is one of those parameters that is designed to differentiate between foreground and background for the one-stage UAV detection during training. We obtained 0.164 focal loss for classification and 0.15 with photometric and geometric augmentation and boundary box augmentation at the end of 100 epochs to train the hybrid model. The regression loss is obtained by using mean absolute error, and the regression loss ends up with 0.99 without augmentation and 0.97 with augmentation. Both classification and regression losses are given in Figure 3.7. The diversity of the aerial scene creates variations in intra-class. These variations affected the classification and regression loss and degraded the scene classification performance. Batch loss (BL) is studied to eliminate the large variations of intra-class features with in-depth features in the batch and batch center of the given UAV. The equation 3.4 represents the batch loss which is given as;

$$BL = \frac{1}{2}\sum_{i=1}^{N}(x_i - B_c)^2, \tag{3.4}$$

where B_c is the center of the corresponding batch and x_i denotes the deep learning features in the batch. We achieved a classification batch loss of 0.163 with simple training and 0.04 with photometric and geometric augmentation. Moreover, the regression batch loss is 0.99 and 0.5, respectively.

Mean average precision (mAP) is the essential evaluation metric to measure a hybrid model UAVs detection performance. The mAP is the mean of all corresponding average precision for a given recall. It defines the model's accuracy. A higher mAP denotes a high precision for the detection of objects. The mAP is defined as follows:

$$mAP = \frac{\sum_{q=1}^{Q} AP_q}{Q} \tag{3.5}$$

As shown in Figure 4.8, we got a mean average precision of 0.658 by using dual streams (day and

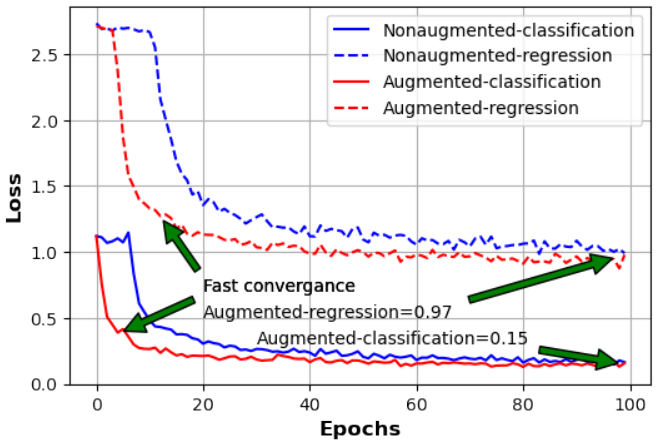


Figure 3.7: Classification and regression Loss using dual stream with the augmented fast convergence and non-augmented classification and regression loss has been shown in this figure.

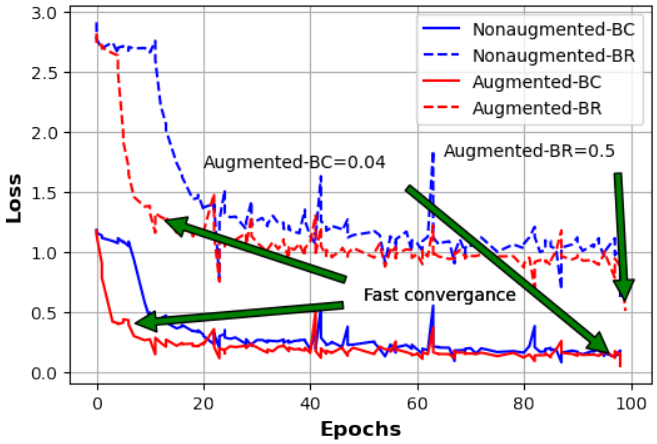


Figure 3.8: In this figure the batch classification (BC) and regression (BR) losses for both augmentation and non-augmentation using boundary box geometric augmentation techniques are shown for the dual stream training.

night vision) of the combination of RGB and IR video frames. We got 0.64 mAP with training the hybrid deep learning model with augmentation. The augmented data training converged very fast as compared to the non-augmented training data. The proposed hybrid deep learning model

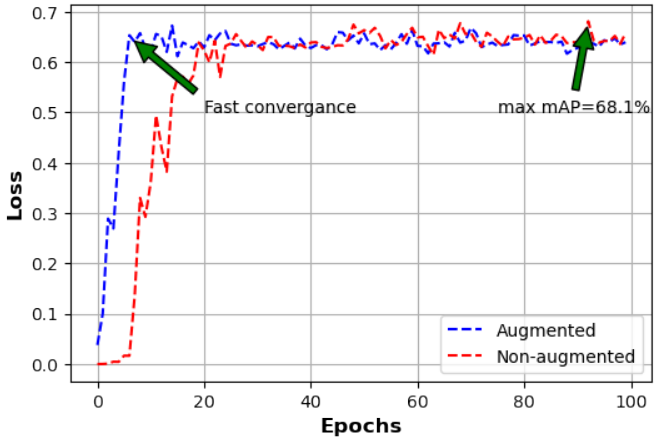


Figure 3.9: Mean average precision (mAP) from the training for both streams with and without photometric and geometric augmentation and bounding box augmentation.

is trained using ImageNet pre-trained weights. We achieved the training and validation loss of 1.156 and 1.57 before photometric and geometric augmentation and 1.13 with 1.5 after boundary box augmentation. Also, faster convergence is a key point to train the model with photometric and geometric augmentation. The training and validation loss is depicted in Figure 3.10, and the validation base losses for the classification and regression with and without batch-wise are given in Figure 3.11. We achieved validation classification loss of 0.39 and 0.4 for validation data. The validation regression loss is 1.17 and 1.11, respectively, which is acceptable for detecting UAVs in videos with high precision.

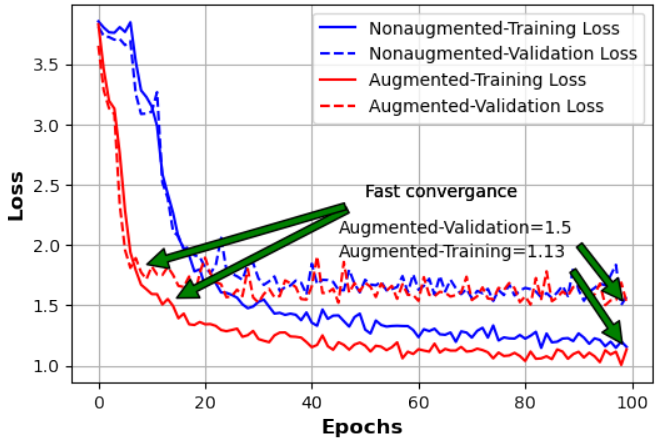


Figure 3.10: Training and validation loss of the hybrid deep learning model with dual stream UAV detection using 100 epochs with and without photometric and geometric augmentation.

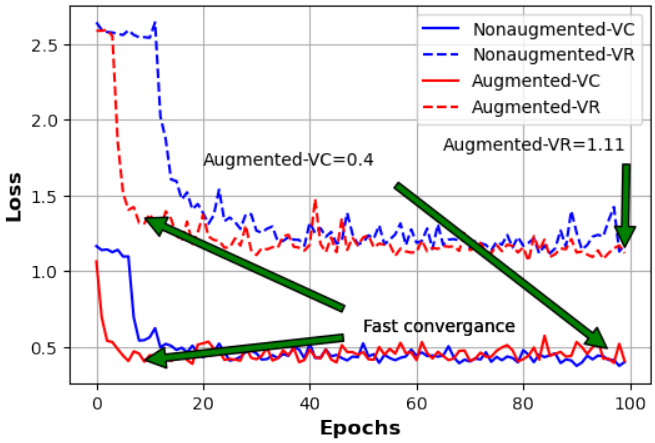


Figure 3.11: Validation classification (VC) and regression (VR) losses of proposed model for both augmented data and non-augmented boundary boxes are shown in this figure.

We used three different videos from diverse backgrounds to check the trained hybrid deep learning model on seen and unseen data. The testing videos are from RGB, IR, and local videos captured of the flying UAV. The trained hybrid model performed very well with RGB videos of the UAV Catch dataset. The detected samples of the RGB video are shown in Figure 3.12. Each video frame contains the detected UAV with the correct label box, and precision is very high even with rotation of different positions of the UAV. The UAVs detection in night vision is a challenging

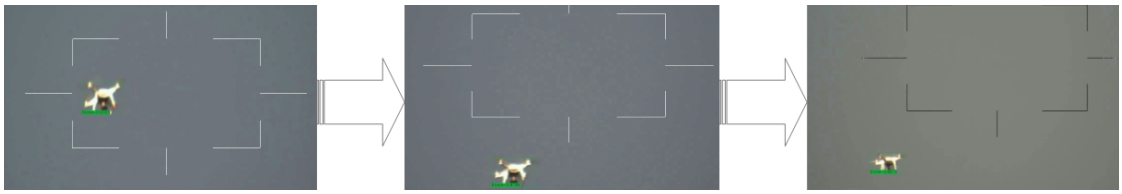


Figure 3.12: UAV detection from an RGB sample of day vision from test data.

problem; we tested the proposed hybrid model for the night vision of the IR stream. Our hybrid model performed best and accurately to identify the UAV and differentiate between UAVs and birds; Figure 3.13 shows the IR-based UAV detection.

Furthermore, it is essential to test on the local data for the real-time application. The proposed



Figure 3.13: UAV detection from infrared (IR) videos of night vision samples from test data.

hybrid deep learning model is used for locally collected flying UAV datasets and achieved the best accuracy, which is shown in Figure 3.14.

Lastly, we are presenting the misclassification of the defined hybrid model. Our hybrid deep



Figure 3.14: UAV detection from real-time videos of on-campus samples from test data.

learning model still needs improvement in terms of precision. The hybrid model sometimes misses the UAV when the background and UAV have the same color. The UAV Catch dataset is used to test other objects with similar or similar UAV shapes detected by the proposed hybrid deep learning model. Similarly, it also misinterpreted the new local dataset background objects. The problem can be eliminated by adding similar data for the training because machine learning needs the same distribution dataset for the correct prediction. The samples of the misclassification are available in Figure 3.15.



Figure 3.15: Miss-classification samples of UAV from different testing videos of RGB, IR videos of both day-night vision of the UAV Catch dataset, and on-campus testing.

3.3 Chapter Summary and Concluding Remarks

In this study, we presented a new hybrid deep learning model for the UAV detection based on a one stream structure for day-night dual visions. The proposed hybrid model highly improves the accuracy of the UAV detection with the RGB and IR videos. To solve the annotation error problem, we propose an algorithm to determine the accurate coordinate dimension and locate the UAVs using a search range for template matching with a brightness threshold and to calculate the drift-corrected displacement between consecutive video frames.

To mitigate the color, UAVs rotations, and axis aligned bounding box problems, we studied photometric and geometric augmentation, which improves model robustness and performance. photometric and geometric augmentation changes the background color, creates clockwise and anti-clockwise rotations and flaps the bounding box along with the UAVs position. Moreover, photometric and geometric augmentation reduces generalization error to play as a regularizer parameter for the model. Similarly, the model's robustness and generalization can be intended by photometric and geometric augmentation for both independent and identically distributed data and out-of-distribution data in a real world problem of UAV detections.

We used the cosine annealing strategy to solve and optimize the training time of the proposed model, achieving significant improvements in training efficiency. The initial layers of the model extract low-level features and require less frequent updates and cosine annealing, adopting a hierarchical feature learning style to prevent frequent updates for convolutional neural networks' initial layers. The higher-level layers focus on high-level feature abstractions, which create very rapid parameter changes during training of the model. To freeze layers at once creates a feature learning problem, and it is necessary to gradually freeze initial layers. Cosine annealing establishes strategic and stable configurations that enhance the stability of the hybrid model. It reduces the computational cost of the training, which accelerates the overall learning process, and freezes model layers gradually to preserve pre-trained EfficientNet valuable features in the early stage of the training model. Cosine annealing, when combined with stochastic depth, enhances the robustness and flexibility of the model by randomly removing layers during the training phase.

Chapter 4

Fusion Flow-enhanced Graph Pooling Residual Networks for UAVs Surveillance in Day and Night Dual Visions

In this chapter we explore **Challenge 2-** with the proposed system model of optical flow-assisted graph pooling residual networks. As we know, identifying UAVs through camera imagery presents a considerable challenge. This arises from UAVs' propensity to integrate inconspicuously with environmental elements, such as structures and foliage, particularly during nocturnal hours when their hues can closely resemble the backdrop. Moreover, during daylight, the variability in illumination conditions further complicates the detection process. In our previous chapter, we explored a deep learning model, leveraging transformations and cosine annealing strategies to reduce classification and regression discrepancies for UAV detection utilizing both RGB and IR imagery. However, the detection efficacy using RGB images is compromised under less-than-ideal lighting conditions, such as during overcast or tempestuous days. On the other hand, while using IR (monochromatic) imagery, discerning UAVs that merge with backgrounds of a similar hue presents its own set of challenges.

We propose optical flow-assisted graph pooling residual networks (OF-GPRN) designed for intricate UAV detection using combined RGB and IR images. While the IR image remains unaffected by light conditions, the RGB image retains vital color data. The proposed OF-GPRN takes advantage of integrating RGB and IR images to optimize contrast, edge definition, color, and texture in each frame. This combined image also mitigates distortions arising from lighting variances, color, and background interference. Relying on this image integration, the OF-GPRN system produces a comprehensive composite frame enriched with features such as fine texture, broad texture, and contrast, which facilitates the extraction of UAV movement patterns.

To isolate the UAV from its background, our proposed OF-GPRN system harnesses the fusion of RGB and IR images, subsequently processed through optical flow (Jia et al., 2021), facilitating the segregation of the UAV from the amalgamated image. This system innovatively augments graph neural networks (GCN) by integrating graph residual split-attention networks (GRSaN) (Li et al., 2021a; Zhou and Chi, 2019; Li et al., 2022c), aiming to optimize the mAP for UAV detection (Zhang et al., 2020). Given the diminutive representation of the extracted UAV within the image, it poses challenges in distinguishing it from other entities, such as avian creatures or aircraft. Specifically, the OF-GPRN model refines the extracted object's contours and ascertains pixel correlations across the pre-processed RGB and IR imagery. This aids in UAV identification and augments predictive accuracy by capitalizing on the expansive feature-learning prowess offered by the feature pyramid during model calibration. Moreover, the main contributions of this chapter are listed as follows:

- The OF-GPRN system is proposed to enable precise UAV detection in day and night dual visions that experience time-varying lighting conditions and high background similarities. The OF-GPRN system develops a fusion of RGB and IR frames, which enhances the quality of output frames by reducing noise and adjusting illumination and color. The OF-GPRN system also extends a new optical flow model to eliminate background and foreground similarity while extracting the UAV's mobility.
- The GRSaN is extended in the OF-GPRN system to stabilize the learning capability, enhance feature learning during training, and reshape the UAV. The OF-GPRN system also uses a Quickshift-based algorithm to represent the adjacency matrix from pixels to nodes. This makes for an accurate graph with clear frame-pixel information.
- We conducted experiments to assess the performance of the proposed OF-GPRN system in UAV detection during both daytime and nighttime conditions. In comparison to preceding residual GCN (ResGCN) object detectors, our enhanced model demonstrates superior performance, achieving a commendable mAP of 87.8% on the stringent RGB-IR combined benchmark UAV catch dataset. This marks a significant improvement over the ResGCN, which obtains an mAP of 69.9%.

This chapter is organized as follows: The proposed OF-GPRN system is presented in Section 4.1. In Section 4.2, we study the system implementation and experimental setup, as well as the performance evaluation. Section 4.3 concludes the chapter Summary and Remarks. The symbols used in the chapter have been listed below in the Table. 4.1.

Table 4.1: Basic Notations

Symbol	Definition
$\mathscr{L}_{(x,y)}$	Modified Laplacian operator.
$L_{(x,y)}$	Modified Laplacian.
L''	Linear differential operator.
F_s, C_s, B_s	fine structure, coarse structure, and base layers.
$w(P,Q)$	Weighting matrix filter with parameters P and Q.
$\mathscr{V}_p$	Visual saliency map.
H, h_n	Concatenated vector, D-dimensional features vector.
W, G_{l+r}	Learnable weighted parameter, graph residual split attention network layer.

4.1 The Proposed OF-GPRN System Model

In this section, we study the proposed OF-GPRN system that learns the feature layers while improving the detection accuracy, which is illustrated in Figure 4.1. The OF-GPRN is developed to retain and recognize the original structure of the complex day and night vision data.

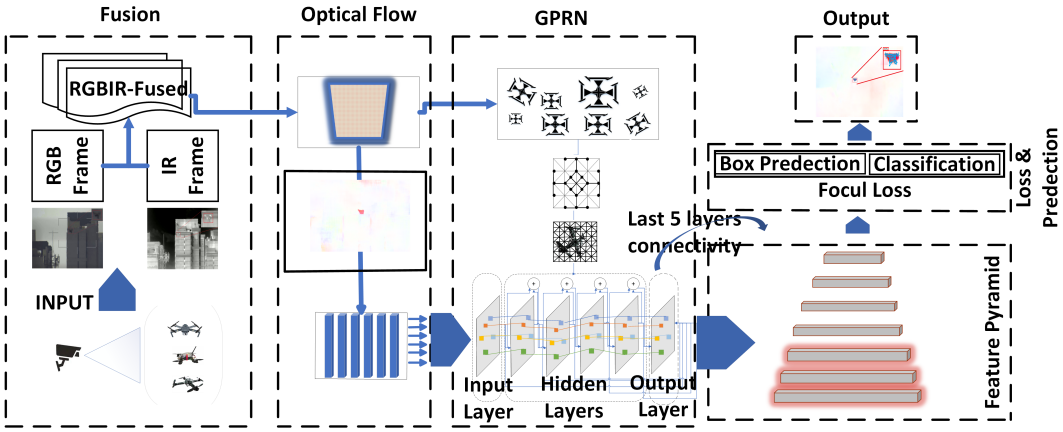


Figure 4.1: The proposed OF-GPRN system architecture for UAV detection and tracking in day and night vision.

4.1.1 RGB-IR Fusion

A fusion module is developed in the proposed OF-GPRN to merge two input frames from RGB and IR sources while enhancing the quality of the output frames by reducing noise and adjusting illumination and color. First, multi-level edge preservation filtering is used to separate the input frames into base layers (B_s), fine-structure (F_s), and coarse-structure (C_s). This enables the extraction of fine texture and large texture features at F_s and C_s layers respectively, while contrast and edge details are preserved in the B_s layer. To retain edge information, weighted mean curvature (W_f) is applied to each input layer, and a Gaussian filter (G_f) is utilized to remove Gaussian noise. The weighting matrix filter $w(P,Q)$ assigns higher weights for center pixels within the square-shaped window, with parameters P and Q. A modified Laplacian operator $\mathscr{L}_{(x,y)}$ is defined to obtain high-quality output frames that capture detailed features, such as edges and contours, from both RGB and IR sources (Yin et al., 2014; Ullah et al., 2021). Thus $\mathscr{L}_{(x,y)}$ is given as

$$\mathscr{L}_{(x,y)} = \sum_{P=-p}^{p} \sum_{Q=-q}^{q} w(P,Q) \left[L_{(x+P,y+Q)} \right]^2 \tag{4.1}$$

Furthermore, the modified Laplacian can be given by $L_{(x,y)} = |2L''_{(x,y)} - L''_{(x-1,y)} - L''_{(x+1,y)}| + |2L''_{(x,y)} - L''_{(x,y-1)} - L''_{(x,y+1)}|$, where L'' is a linear differential operator that approximates the second derivative at the F_s layer, i.e., $L''_{F_{s(x,y)}} = \{\frac{\delta^2 F_s}{\delta x^2} + \frac{\delta^2 F_s}{\delta y^2}\}$. Likewise, $L''_{C_s(x,y)}$ at the C_s layer can also be obtained. In the proposed OF-GPRN, a pulse-coupled neural network with parameter adaptation is used to fuse the F_s and C_s layers while determining the optimal number of features in each layer (Yin et al., 2019). Specifically, the network evaluates the edge features of the respective layers, with priority given to those extracted from $\mathscr{L}_{F_{s(x,y)}}$ if they are more prominent. If the edge features of $\mathscr{L}_{C_{s(x,y)}}$ are more prominent, the pixel features extracted from this layer are added to the fused frame.

The fusion of the B_s layers in the RGB and IR frames combines the contrast information from an IR frame with the texture information from an RGB frame. In particular, fusing the B_s layers in RGB and IR frames suffers from each pixel contrast and texture information composition (Yan et al., 2021). Therefore, a visual saliency map , denoted by $\mathscr{V}_p$, is constructed based on calculating the intensity value of the pixels, which examines the difference between each pixel and all of its neighbors to generate a saliency value (4.2). As a result, $\mathscr{V}_p$ preserves the contrast and texture features of the RGB or IR frame, improving the quality of the fusion. In (4.2), n represents a specific pixel intensity value within the range of 0 to 255 in the frames $B_s^{(RGB,IR)}$ (denoted as N), and $\mathscr{I}_n$ is the number of pixels with similar intensity to $n \; \forall \; I = \{i|(x_{n_i}, y_{n_i}) = n, 1 \leq i \leq N\}$ in which i is the individual pixel index. The proposed system uses a feature scaling normalization function $S(.)$ to make sure that the frame features fall in the same range. This is due to the fact that different scales in (4.2) affect many quantitative pixel features.

$$\mathscr{V}_p = \sum_{p=x,y}^{x_n,y_n} \mathscr{I}_n \left| S\left(B_{s_p}^{(RGB,IR)}\right) - S\left(B_{s_{x_n,y_n}}^{(RGB,IR)}\right) \right|, \tag{4.2}$$

where x and y are spatial pixel coordinates, and x_n and y_n represent specific pixel coordinates.

For the fusion of the B_s layer, we formulate (4.3) to merge the saliency maps in the RGB and IR frames, which are represented by $\mathscr{V}_p^{RGB}$ and $\mathscr{V}_p^{IR}$, respectively.

$$BASE_{fusion} = \frac{\alpha + \beta}{2}, \tag{4.3}$$

where $\alpha = (\mathscr{V}_p^{IR} B_s^{\mathrm{IR}} + (1 - \mathscr{V}_p^{IR}) B_s^{\mathrm{RGB}})$ and $\beta = (\mathscr{V}_p^{RGB} B_s^{\mathrm{RGB}} + (1 - \mathscr{V}_p^{RGB}) B_s^{\mathrm{IR}})$. Based on the output of the parameters-adaptive pulse coupled neural network and $BASE_{fusion}$, the fusion of the RGB and IR frames can be obtained by applying the inverse multi-level edge preservation filtering.

To monitor the UAV's movement, we used optical flow BRAFT (Jia et al., 2021) to process the frames, which relies on merged frames. an effective approximation of the actual physical motion being projected onto the fused frame, which provides a concise representation of the parts of the frame that are in motion. Integrating spatio-temporal information helps with background elimination. Once the videos are processed, the best next move for the mobile UAV is determined, and routine execution initiates each new iteration of the model.

4.1.2 Graph Residual Split Attention Network (GRSaN)

The construction of graph nodes and edges for input images in the proposed OF-GPRN model depends on using region adjacency graphs. Superpixel segmentation approaches, such as SLIC, Quickshift, and Felzenszwalb, are used to precisely split the frames into regions, which then function as nodes in a system. Subsequently, each of these regions connects together on the basis of their adjacency, which leads to the formation of the edges of the graph. The decision of the superpixel segmentation technique plays a role in determining the precision of a generated graph. These nodes and edges are used as inputs for the GCNs model. GCNs are graph-based architectures based on graph nodes and edges $G = (N, E)$. Instead of using conventional convolutional filters, GCNs use graph convolutional filters in each layer of unordered nodes N with edges E, and aside from that, GCNs are just like CNNs. Stacks of pointwise nonlinearities in GCNs serve as the building blocks of filters, while stability and permutation equivariance of GCN architectures with good performance are attributed to the graph characteristics (Ruiz et al., 2021; Li et al., 2021a). UAV pixels in the frame are represented as nodes $N = [n_1, ..., n_k] \in \mathbb{R}$, with edges $E = [e_1, ..., e_k] \in \mathbb{R}$ defining the relationship between the i-th and j-th UAV pixels in the order pair $k = (i, j)$. The vector $H = [h_{n_1}, h_{n_2}, ..., h_{n_k}]^T \in \mathbb{R}$ concatenates the feature vector $h_n \in \mathbb{R}$ with D-dimensional features of n nodes. Here's how the information from the input layer (G_l) which is $ConvOper(G_l, W_l)$ and

added to and changed in the output layers.

$$\begin{aligned} G_{l+out} &= ConvOper(G_{l+out}, W_{l+out}) \\ &+\tau(G_l + G_{l+1} + ... + G_{l+out-1}) \end{aligned} \tag{4.4}$$

$W = [W_1, W_2, ..., W_{out}]$ is the learnable weighted parameter of the n layers for node aggregation and updating the graph function to compile neighborhood pixel information (Li et al., 2021a).

Each G_{l+r} represents the graph residual split attention network layer of the graph residual network (Zhang et al., 2020) as shown in Figure 4.2. The output of G_{l+r} is given in equation 4.5.

$$G_{out} = \sum_{r=1}^{R} (G_{l+r}^{Conc}) + \tau(G_l), r = 1,2,3,4,\ldots R\} \tag{4.5}$$

G_l is the input of the graph residual split attention network layer, G_{out} is the output, and τ is the strided graph convolution or combined graph convolution with max pooling. If the dimensions of G_{out} and G_l equal, then τ replaced by the identity matrix $(\mathbf{I})$ (Kim et al., 2017b). The GCN main route output is scaled using τ as a linear projection with the previous input. τ is scaled with input using a strided graph convolution or a combined graph convolution linear filter with max pooling. As a result, the number of parameters for ResGCN remains constant rather than increasing, as it does for plain GCN or ResGCN without τ. A simple ResGCN without a projection matrix block can add an input channel to a GCN output; however, as the number of layers increased, performance decreased significantly due to shortcut path accumulation. Furthermore, if the input and output have the same dimension, then $\mathbf{I}$ can reduce computational complexity and have the same effect.

Where G_{l+r}^{Conc} represents the concatenation of the cardinality groups denoted by (k) for each set of hyperparameters (R). To have a better understanding of each expression, we listed a more in-depth comprehension of each term:

- G_{l+r}^{Conc}: This notation refers to a specific data structure that results from concatenating multiple groups, where each group corresponds to a different choice of hyperparameters (R).
- $\sum_{r=1}^{R} (G_{l+r}^{Conc})$: This expression represents the sum of these concatenated groups. Moreover, it is adding together the information contained in all the different groups. The result is a comprehensive dataset represented as $G_1 + G_2 + \ldots + G_k$, where each $G^k \in R^{N \times D}$ corresponds to a specific combination of nodes and the cardinality of the set k.
- $G^{Conc} = H_G(G^1, G^2, \ldots, G^k)$: We define a function H_G that takes individual groups $G^1, G^2, \ldots, G^k$ as inputs and concatenates their gradients (features). This step ensures that we capture information from all the vertices in each group.

$$G^k = \sum_{j=1}^{k} F((G_j) + (w_j^{agg})) + w_j^{update} \tag{4.6}$$

where the updated and aggregated learnable parameters are w_j^{agg} and w_j^{update}. Where $F(.)$ is the aggregation function. w_j^{agg} compiles information from vertices in the same cardinal k's neighborhood, whereas w_j^{update} applies a non-linear function to the aggregated information to compute new vertex representations in cardinal k. The G_{out} is processed by global max pooling, followed by batch normalization and ReLU to stabilize the input by softmax and 1x1 convolution, and then transferred for the next layer, as shown in Fig. 4.2.

4.1.3 Graph-Pooling Feature Pyramid Network Mapping

The last layers of the GCN extract high-level features of the input. We selected the last 5 layers for the graph feature pyramid network to map features between GCN and the pooling feature pyramid (PFP) (He et al., 2015; Lin et al., 2017). We update the last 5 layers of the GCN with a feature pyramid network (Zhao et al., 2021), which has the ability of the superpixel hierarchy to make recursively larger groups of pixels from the smaller features of the last layers (high-level features) and a similarity measure (Pont-Tuset et al., 2017). The superpixel hierarchy matches the graph layers, and when moving from one layer of the residual attention GCN last layer to the next, the number of nodes decreases by a factor of 4. Contextual and hierarchical edges are used in different ways in the layer that connects the ancestor and descendant layers, which are called superpixels. Hierarchical edges connect semantic gaps, and contextual edges spread information about the context of the different levels in each layer. The node features used are the same for both hierarchical and contextual layers, but the edges are different for both. Of the last 5 layers, the first and last two are contextual, while the middle layers are hierarchical. Moreover, the learning parameters are different for both and are not shareable. The mapping from GCN to PFP is necessary to transfer multi-scale features and ensure compatibility with the PFP input. Each input from the GCN layer is processed with a stride of 2, which preserves the hierarchical feature representation and maintains effective feature propagation across layers. An upsampling factor of 2 is applied to features with a higher resolution. Each lateral link combines feature maps of the top-down pathways that are the same size in space. Convolutions are performed on the top-down route feature maps to lower the channel dimensions, and the feature maps from both pathways (input: N/4, N/16, N/64, and N/256; + output: N, N/4, N/16, N/64, and N/256) are combined using element-wise addition. Each combined map is given a 3 x 3 convolution with a factor of 2 to get the final forecast for each layer. For the UAV's position, generative localization of bounding boxes makes a single box with the highest score.

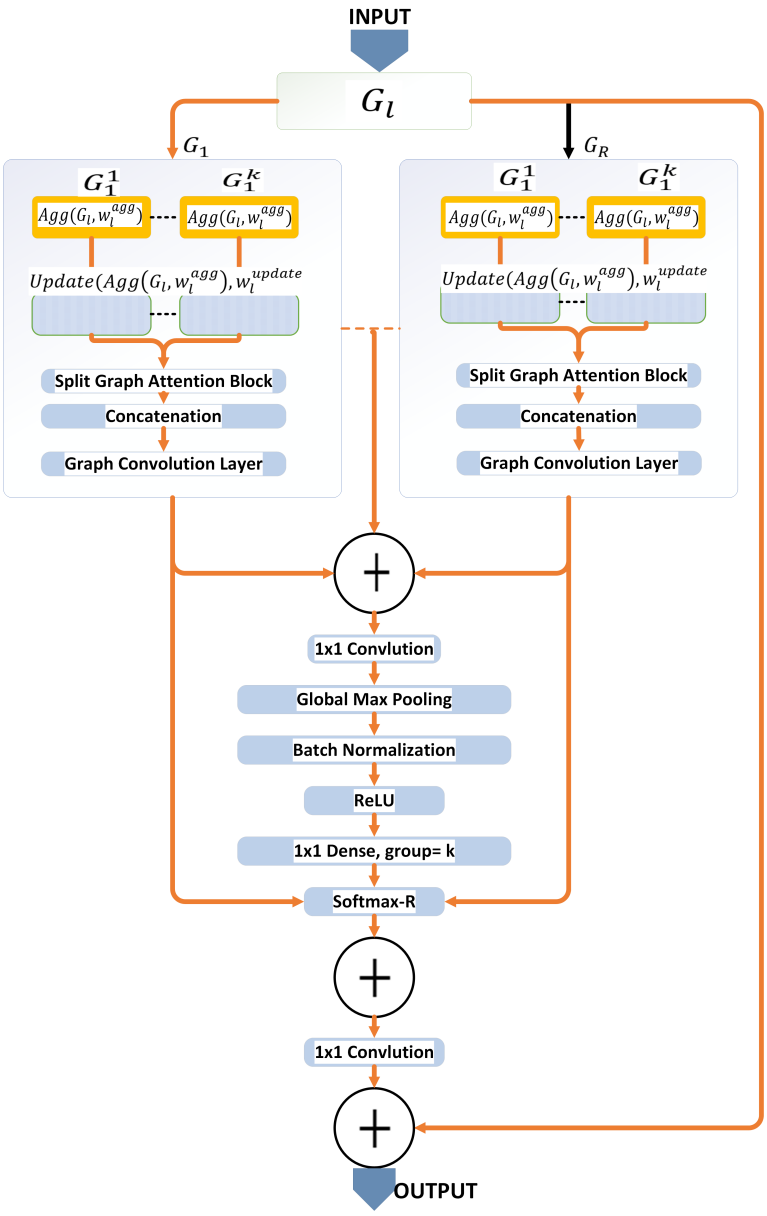


Figure 4.2: In a deep overview of the split graph attention is shown here, which represents each layer of the graph convolutional network.

4.1.4 Loss Function

In the case of one-stage detection, focal loss is specifically tailored to meet the needs of the user. In the suggested training model, an imbalance between UAVs in the foreground and those in the background could be fixed to put less weight on making accurate predictions. It is common practice to use cross-entropy as a loss function because of its high level of accuracy in comparing the approximation models.

$$Cross\ Entropy(_{p,t}) = \begin{cases} -\lambda 1_t\ log(p), \text{ if } t = 1 \\ -\lambda 1_t\ log(1-p), \text{ otherwise} \\ \text{When, } t \in \{\pm 1\}, p \in [0,1] \end{cases} \tag{4.7}$$

In (4.7), t is the value of the UAV detection target and p is a probabilistic estimate of that target value based on the probability distribution. Where $\lambda 1_t$ are the balanced parameters for positive and negative examples; however, it cannot discriminate between simple and challenging cases. The downweight approach requires the modulated focal loss factor $(1-p_t)^{\gamma}$ for numerical stability.

However, when training naively with (4.7), the classifier is unable to discriminate between the more accurate candidate and the loose counterpart, resulting in an unanticipated learning scenario as shown in Fig. 4.3. Because the candidate boxes with more precise locations are suppressed with non-maximum-suppression procedures, this may have a negative impact on performance. As shown in Fig. 4.3, the consistent cross-entropy loss function is a dynamically scaled cross-

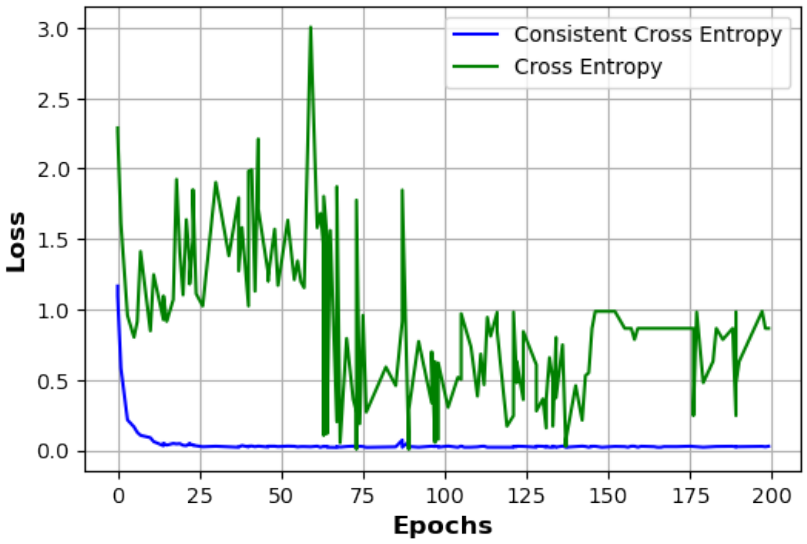


Figure 4.3: The scaling factor's impact on cross-entropy to automatically downweight loose samples during training.

entropy loss with the scaling factor determined by the overlap between the current bounding box and the target ground-truth item (Sun et al., 2019). This scaling factor, intuitively, automatically

downweights the contribution of loose samples during training, allowing the model to concentrate on more accurate predictions. The consistent cross-entropy loss function may help train our model to better identify which prediction is the best among numerous clustered choices. So, modulating factors are added to (4.7) using a consistent cross-entropy loss function by Sun et al. (2019) to accommodate localization quality, and the more precise targets are augmented to reflect it. The (4.7) updated form is shown in (4.8).

$$Cross\ Entropy_{(p,t)} = \begin{cases} [-\lambda 1_t + \lambda 2(o_k - \lambda 1)z_k]\ log(p), \text{ if } t = 1 \\ [-\lambda 1_t + \lambda 2(o_k - \lambda 1)z_k]\ log(1-p), \text{ otherwise} \\ \text{When, } t \in \{\pm 1\}, p \in [0,1] \end{cases} \tag{4.8}$$

The $z_k := 1(if o_k > \alpha, k = 1,...,L)$ represents the candidate box of the targets using IoU overlap for the predicted bounding box and o_k shows IoU overlapping of the predicted and ground truth bounding boxes. Here k is the location of the UAV in the frames. Frames with IoU overlap more than may use the modifying factor in (4.8) for more favorable examples, which increases the modifying factor and the loss in proportion to the overlap with ground truth targets. As a result of using cross-entropy, the consistent cross-entropy loss function prioritizes cases with bigger IoU overlaps.

The (4.8) is updated in the focal loss as:

$$Focal\ Loss(_{p,t}) = -\alpha_t\ (1-p_t)^{\gamma}\ log(p_t) \tag{4.9}$$

The focal loss down-weighted tuning process is dependent on the γ, and it varies from 0 to 2 to adjust the rate of easy examples. If γ is 0, then the focal loss is equal to the cross-entropy. The UAV detection scenarios increase the γ up to 2 to obtain the best training result. Moreover, during experimentation, we systematically varied γ and observed its impact on the performance of the model. Higher γ led to better detection rates for challenging scenarios, such as low contrast or occlusion cases. However, excessively high values might lead to overemphasis on hard examples, potentially causing instability.

4.2 Experiments and Performance Analysis

4.2.1 Experimental Training

Extensive experiments on the benchmark UAV catch detection dataset were conducted to evaluate the efficacy of OF-GPRN in enhancing the learning performance for UAV detection with day and night vision cameras. The OF-GPRN model is developed in TensorFlow and trained on a workstation with a GeForce RT 3060. The training procedure used 4 batch sizes. Compared to other optimizers, the Adam optimizer is preferred because of its rapid and intuitive convergence

on the best solution. The learning rate decays to 10^{-4} and 10^{-6} when the value of $\beta 1$ is set to 0.9 and when the value of $\beta 2$ is set to 0.999.

In addition, the model is trained for 45 hours and 200 epochs. To build the region adjacency graph and decrease the input size, the frames are converted to super pixels using the algorithms SLIC (Achanta et al., 2012), Quickshift (Vedaldi and Soatto, 2008) and Felzenszwalb (Felzenszwalb and Huttenlocher, 2004) for OF-GPRN training. Progressive focal loss is used to prevent the unanticipated learning scenario and to discriminate between the more accurate candidate and the loose counterpart. textcolorblue The comparative performance of UAV detection across various models is shown in Table 4.2. The effectiveness of detection demonstrates substantial variation among models, with the proposed model obtaining the highest performance. The assessed approaches consist of RGBIR-ResGCN, Fusion-ResGCN, OF-ResGCN, and the proposed OF-GPRN. The RGBIR-ResGCN model obtained an mAP of 35.1% but suffered from a comparatively high loss of 4.253, suggesting that its performance is unstable. Fusion-ResGCN demonstrated an enhanced mAP of 55.5%. However, there is a 2.147 increase in loss value along with this improvement, primarily as a result of difficulties maintaining background features, which reduces stability. The optical flow-based model OF-ResGCN achieved an mAP of 69.9%. A loss of 2.042, however, indicates that the model's performance is unsatisfactory. In contrast, the proposed OF-GPRN proved significant improvements, with a remarkable mAP of 87.8% with a small loss of 0.026. This model showed higher efficiency than other models and also maintained consistent accuracy in its predictions. The significant increase in mAP, along with minimal loss, indicates that OF-GPRN has the potential to be a very effective method for the given task of UAV detection.

Table 4.2: Experimental Results with Stability and Model Performance

Method	**mAP (%)**	**Loss**	**Comments**
RGBIR-ResGCN	35.1	4.253	Unstable
Fusion-ResGCN	55.5	2.147	High loss due to background
OF-ResGCN	69.9	2.042	Lower performance
OF-GPRN (Proposed)	**87.8**	**0.026**	Stable with improved mAP

4.2.2 Datasets Pre-processing

In this study, we employ freely accessible anti-UAV capture video datasets (Jiang et al., 2021). There are a total of 320 clips here, 160 of which are HD videos shot in both standard definition (SD) and high definition (HD) (RGB and IR), which have different variations like backgrounds

(cloud, building, mountain, and sea), fast movement, out of focus, and size variations from small to large. The UAVs seen in each video come in a range of sizes, from large to small. We selected UAVs with a size range of 300mm to 1200mm to train and validate our proposed model. Moreover, the UAVs cruise at speeds between 50 and 100 miles per hour and occasionally stop altogether. One hundred validation videos from each stream were used for model training. All except 80 of the videos are used to train the model, while the remaining 20 are used for testing and validation.

There are a wide variety of variables in the background that may be visible in the video clips, including day and night, lighting conditions, cloudy and clear skies, buildings, and varied degrees of occlusion, as shown in Fig. 4.4. These background variations make it difficult for the detection algorithms to identify the movable object due to low contrast, weak edge details, color similarities, and texture information. The dataset frames are preprocessed through RGB-IR fusion and an optical flow algorithm to make the movable objects visible.

4.2.3 Effect of Optical Fusion

The conversion of RGB or IR to optical flow is seen in Fig. 4.4. When compared side by side, all of these frames have distinct levels of performance. A closer inspection reveals obvious artifacts, blurriness, and distinctions in the results of the three columns. When compared to fusion-generated frames, we observe that optical flow fails to inject as many of the bright feature characteristics from the RGB (first column) and IR (second column) frames. The fusion-generated RGB and IR (third column) frames support the optical flow to detect the moveable UAV. Without fusion, the results of optical flow show that the moveable UAV is unidentified due to its high background similarities. Therefore, the spatial properties of the source frames are enhanced in the fusion plus optical flow output frames, which are also free of artifacts, clearer, include more structural details, and have a higher overall visual quality.

4.2.4 Effect of Region Adjacency Graphs

Training the GCN model relies on accurate region adjacency graphs. We can see that the pixel segmentation before the optical, as shown in Fig. 4.5 is very complicated compared to the after optical flow. The optical flow for background removal significantly reduces the segmentation of pixels for the region adjacency matrix up to 30%. Moreover, if the appropriate pixel segmentation procedure is used, the resulting region adjacency graph would be accurate. We used three different superpixel segmentation techniques (SLIC, Felzenszwalb, and Quickshift) to significantly minimize the size of the nodes and edges in the matrix to train the OF-GPRN, as shown in Fig. 4.5. Each of the algorithms identifies regions with similar visual properties. Each frame's features and their associated graph are generated by the adjacency matrix, which also provides concise information regarding the frame's pixels.

From our experience training the OF-GPRN, we can say that the SLIC algorithm is both fast and space-efficient, and it is able to successfully segment in terms of color boundaries without having

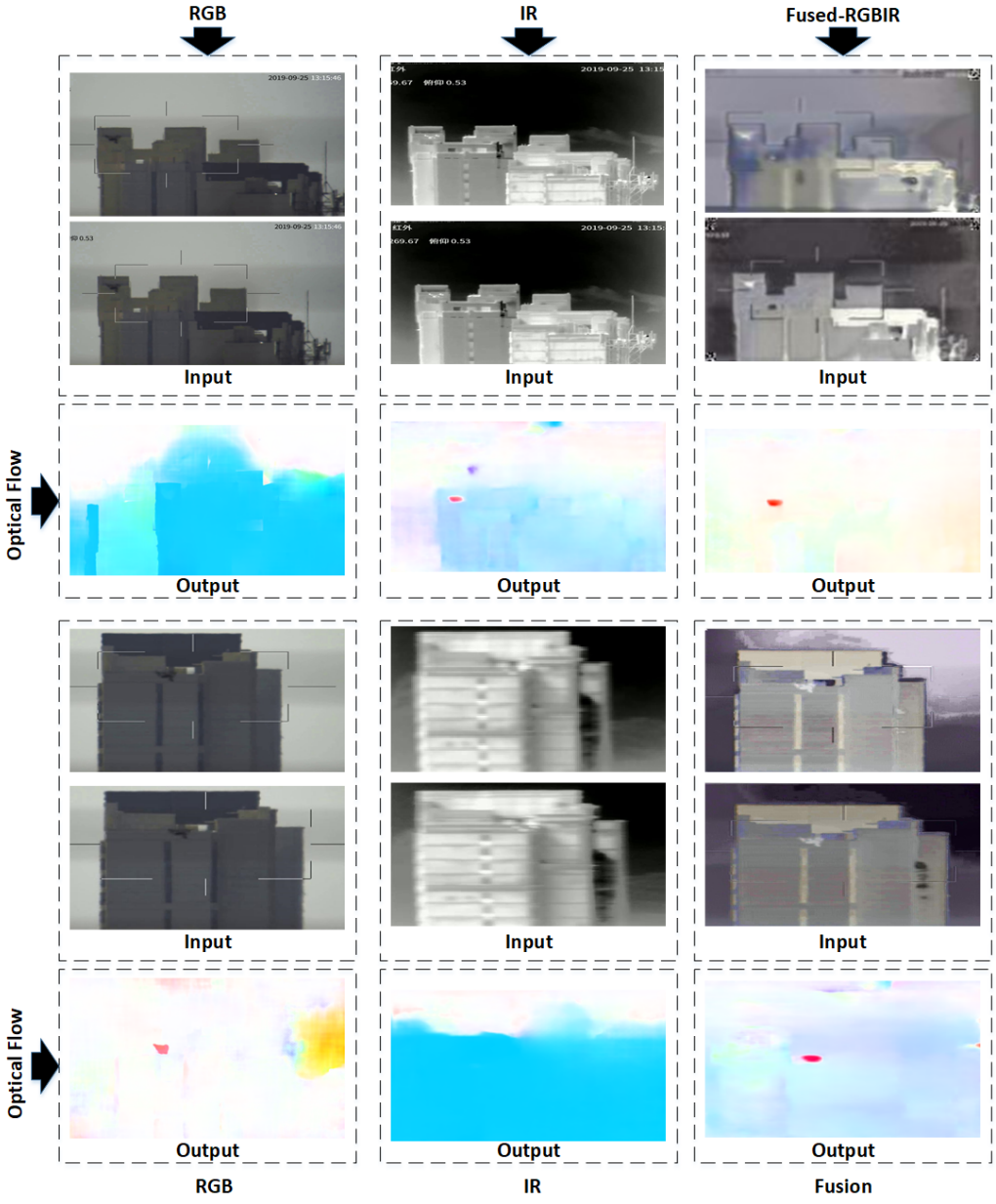


Figure 4.4: The columns on the far right are RGB frames with object optical flow. The middle column shows the IR frames and optical flow. In contrast, the most right-hand column combines RGB-IR and then displays the effect of UAV optical flow.

to remove the background. However, it recorded the pixels in the background, which makes it less accurate. Additionally, OF-GPRN training using the Felzenszwalb algorithm performs less well due to contrast-based training when it comes to loss minimization. While the Quickshift method is used for the adjacency matrix to achieve OF-GPRN-based promising results of loss 0.026, as shown in Fig. 4.6. Quickshift improves in this regard since it uses hierarchical segmentation computation to separate the image into visually distinct parts. The Quickshift algorithm is used for the proposed model due to its high performance. We achieved the optimal loss of all three algorithms with the parameters mentioned in the Table. 4.3.

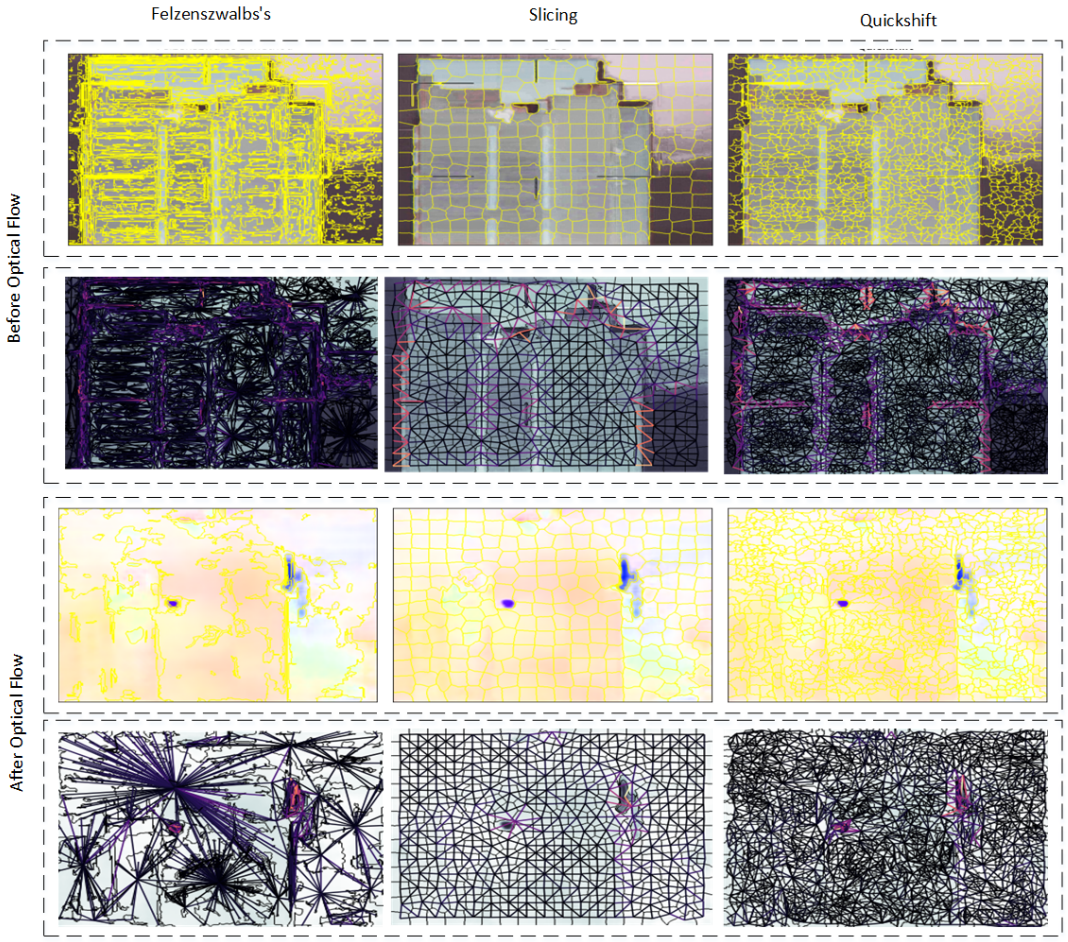


Figure 4.5: The region adjacency graphs before optical flow and after optical flow have three different segment approaches: The leftmost column is the Felzenszwalb algorithm performance, the middle column is the SLIC algorithm results, and the rightmost column is the Quickshift algorithm segmentation for the adjacency matrix.

We use the Quickshift algorithm for exceptional performance after applying optical flow, a method that successfully removes the background. The decision to strategically employ the Quickshift algorithm is based on its notable efficiency in comparison to SLIC and Felzenzswalb. The Quickshift method works better than SLIC and Felzenzswalb in terms of loss efficiency and accurate capture of important features in the image, as shown in Fig. 4.6.

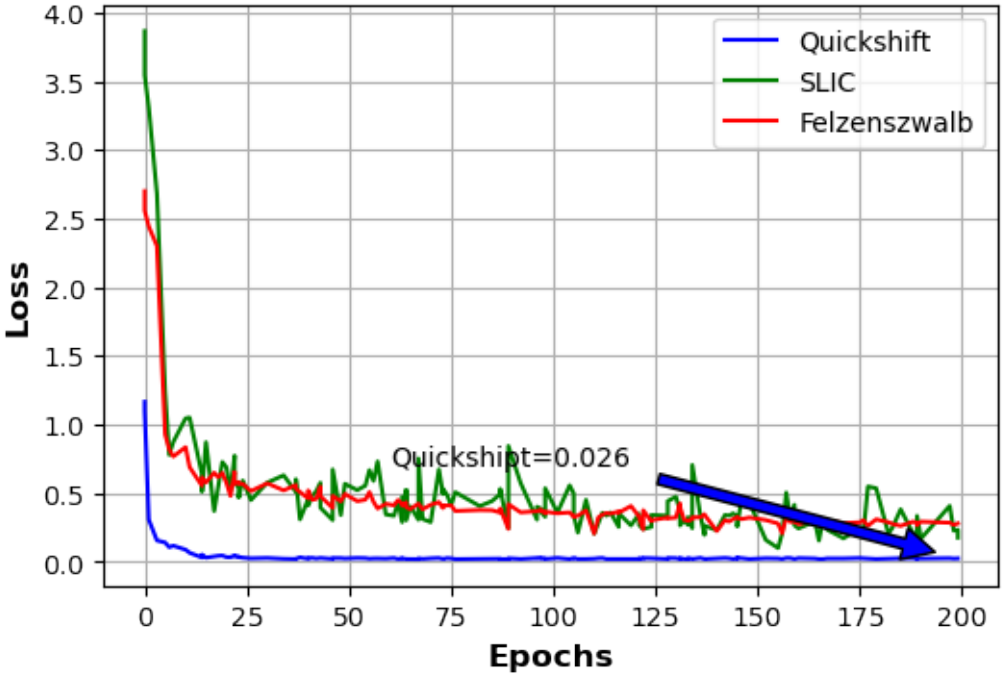


Figure 4.6: The training losses of the OF-GPRN (based on the Split Attention Network) with three different segment approaches for region adjacency graph algorithms after fusion plus optical flow.

Table 4.3: Summary of Segmentation Methods

RAG Method	P_1	P_2	P_3	**Loss**
SLIC	Number of Segments = 100	Compactness = 25	Sigma = 0.3	0.184
	Number of Segments = 1000	Compactness = 10	Sigma = 0.7	0.235
	Number of Segments = 250	Compactness = 20	Sigma = 0.5	0.176
Felzenszwalb	Scale = 40	Sigma = 8	Min size = 0.9	0.322
	Scale = 100	Sigma = 10	Min size = 0.7	0.801
	Scale = 50	Sigma = 10	Min size = 0.5	0.281
Quickshift	Kernel size = 24	Max Dist = 8	Ratio = 0.8	0.158
	Kernel size = 6	Max Dist = 4	Ratio = 0.2	0.059

(continued on next page)

RAG Method	P_1	P_2	P_3	Loss
	Kernel size = 3	Max Dist = 6	Ratio = 0.5	**0.026**

4.2.5 Residual Split Attention Network Effect

Extensive studies, beyond the results of the Quickshift algorithm, are done to illustrate the efficacy of the Split Attention Network in enhancing the learning performance ofGCN with deep architectures. In Fig. 4.7, we show how well, without overfitting, the proposed OF-GPRN (based on the Split Attention Network) with a loss of 0.026 with a learning rate 10^{-6} performs compared to the other ResGCN and observe that the proposed OF-GPRN performs best compared to the RGBIR-ResGCN, Fusion-ResGCN, and OF-ResGCN. The RGBIR-ResGCN has very high instability for the learning rate 10^{-4} and low learning variations during training with a learning rate of 10^{-6} but has a high loss of 4.253. Even Fusion-ResGCN has a very high loss due to background existence in the frame, which leads to a loss of 2.147. Moreover, OF-ResGCN has a lower loss than RGBIR-ResGCN and Fusion-ResGCN; however, it is still unstable, and the loss ends up at 2.042.

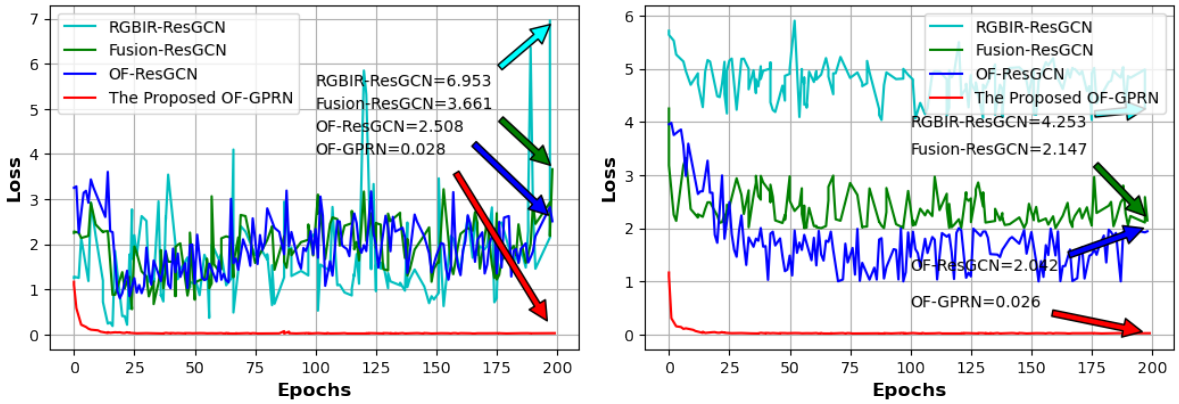


Figure 4.7: The training loss of the four different graph residual networks is: The cyan line represents the ResGCN with RGB-IR, the green line shows the fusion-based ResGCN, the blue line represents the optical flow fusion ResGCN, and the red line represents the proposed optical flow-assisted graph pooling residual network (OF-GPRN). Moreover, (a) is trained with a learning rate of 10^{-4} and (b) is trained with a learning rate of 10^{-6}.

Fig. 4.9 shows both the optimal training loss score of 0.026 and the actual validation loss of 0.331 and the testing loss of 0.35. Moreover, Fig. 4.8 presents the results of the qualitative mAP evaluation with learning rates of 10^{-4} and 10^{-6} and makes a comparison between the OF-GPRN method

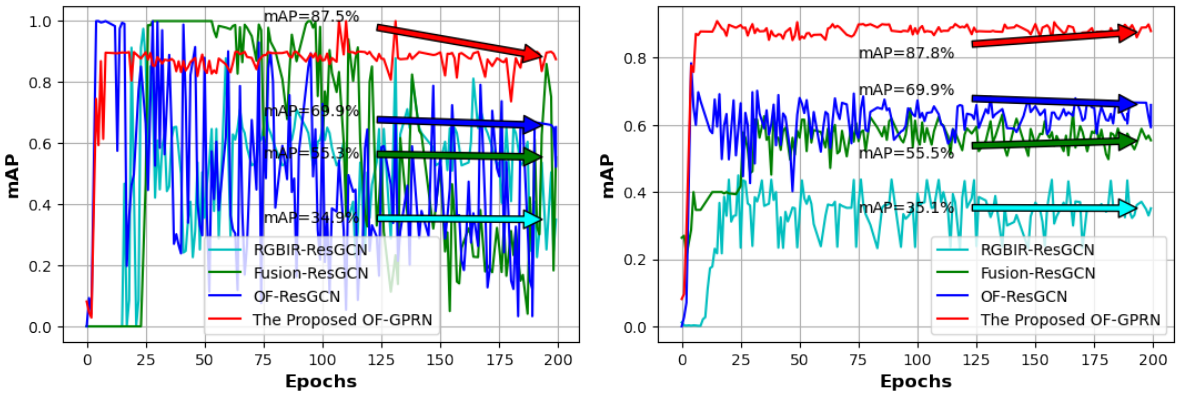


Figure 4.8: The mean average precision (mAP) performance of the proposed OF-GPRN model vs. the RGBIR ResGCN, Fusion ResGCN, and optical flow ResGCN. Moreover, (a) is trained with a learning rate of 10^{-4} and (b) is trained with a learning rate of 10^{-6}.

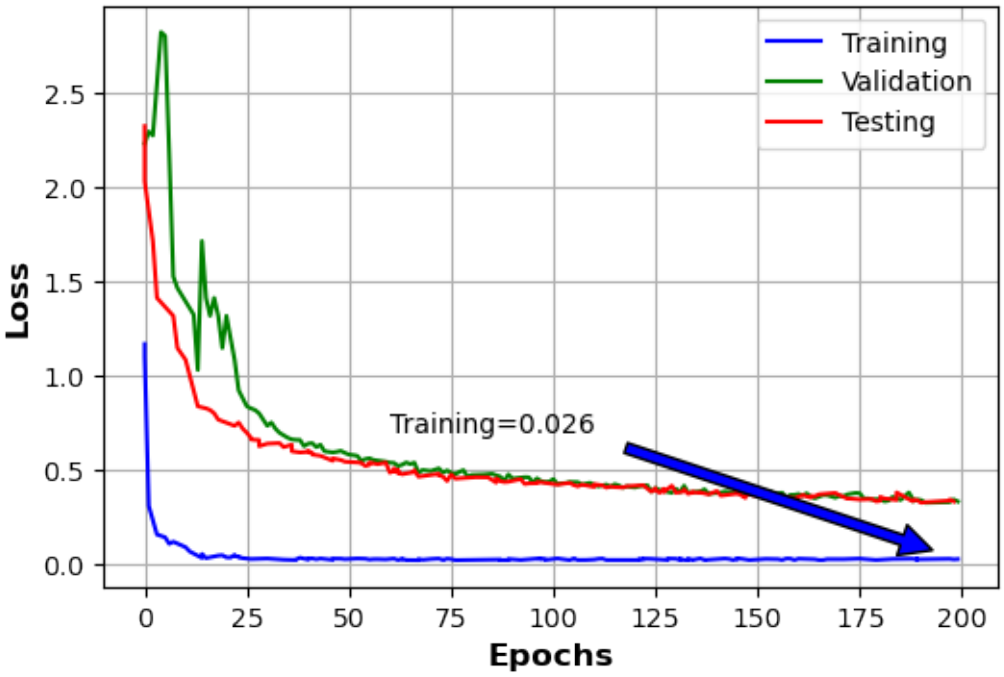


Figure 4.9: The proposed OF-GPRN model performance in the form of training, validation, and testing losses uses Quickshift for the region adjacency graphs. Moreover, advanced focal loss is used to train the proposed model.

and the RGBIR-ResGCN, Fusion-ResGCN, and OF-ResGCN. The OF-GPRN model that has been proposed offers several benefits when it comes to the detection of UAVs. It is clear to observe that the proposed OF-GPRN has a higher mAP with stability during training for different UAV sizes compared to the RGBIR-ResGCN, Fusion-ResGCN, and OF-ResGCN. Moreover, the proposed OF-GPRN achieved 87.8% more than RGBIR-ResGCN, Fusion-ResGCN, and OF-ResGC, which only reach 35.1%, 55.5%, and 69.9% and are unstable while the model is being trained.

As shown in the results of the experiments in Table 4.4 clearly shows that the mAP of the proposed OF-GPRN is much better when it is made up of RGBIR-fused optical flow data. Our proposed OF-GPRN model achieves a highly remarkable mAP of 87.8%. The OF-GPRN model outperforms previous models, such as the Hybrid-DL, which obtained an mAP of 68.1%, and the EfficientDet model, which achieved 67.3% mAP. Importantly, our model shows higher accuracy compared to the DETR model from anti-UAV (Carion et al., 2020), which attained an mAP of 83.2%. The anti-UAV DETR model, EfficientDet, and Hybrid-DL with the Fusion method all have different performance metrics, as shown in the Table. 4.4. Anti-UAV DETR model has a high precision of 87.22%, whereas EfficientDet has a similar but slightly lower precision of 71.11%. Hybrid-DL lies somewhere in the middle, with a precision of 73.55%. In terms of recall, the anti-UAV DETR model is at 79.61%, followed by Hybrid-DL at 67.14% and EfficientDet at 65.33%. The F1 score shows that EfficientDet scored 68.10%, Hybrid-DL scored 70.12%, and anti-UAV DETR model scored 83.24%.

EfficientDet and Hybrid-DL come in second and third, with the anti-UAV DETR model achieving the highest mAP at 83.2% and 67.3%, respectively. In terms of computing efficiency, EfficientDet has the lowest inference time of 102 ms, followed by Hybrid-DL at 107 ms and anti-UAV DETR model at 140 ms. Despite the longer inference time, the proposed model OF-GPRN, outperforms its competitors in precision 91.36%, recall 89.52%, and F1 score 90.43%, indicating a favorable trade-off between performance and computational cost. The proposed OF-GPRN model has a slightly longer inference time of 250 ms but surpasses its competitors in terms of precision, proving its superiority in the application of UAV detection.

As shown in the results of the experiments in Table 4.4 clearly shows that the mAP of the proposed OF-GPRN is much better when it is made up of RGBIR-fused optical flow data. Our proposed OF-GPRN model achieves a highly remarkable mAP of 87.8%. The OF-GPRN model outperforms previous models, such as the Hybrid-DL, which obtained an mAP of 68.1%, and the EfficientDet model, which achieved 67.3% mAP. Importantly, our model shows higher accuracy compared to the anti-UAV DETR model, which attained an mAP of 83.2%.

Upon analyzing the training performance, our proposed approach shows an incredibly small training loss of 0.026. In comparison, the hybrid-DL model has a training loss of 1.13, the EfficientDet model has a loss of 1.20, and the anti-UAV DETR model has a training loss of 0.11. The significant decrease in training loss for our model shows its high efficiency and efficacy in acquiring

features from the data. The results highlight the importance of using RGBIR-fused optical flow data to improve the overall performance of the proposed model compared to other methods.

Table 4.4: Performance Comparison of Detection Models with GPUs

Method	**Precision**	**Recall**	**F1 Score**	**mAP (%)**	**Loss**	**Inference**
Anti-UAV (DETR) (Isaac-Medina et al., 2021b) + Fusion	87.22	79.61	83.24	83.2	0.11	140ms
EfficientDet (Tan et al., 2020) + Fusion	71.11	65.33	68.10	67.3	1.20	**102ms**
Hybrid-DL (Noor et al., 2021) + Fusion	73.55	67.14	70.12	68.1	1.13	107ms
OF-GPRN (Proposed)	**91.36**	**89.52**	**90.43**	**87.5**	**0.026**	250ms

Upon analyzing the training performance, our proposed approach shows an incredibly small training loss of 0.026. In comparison, the hybrid-DL model has a training loss of 1.13, the EfficientDet model has a loss of 1.20, and the anti-UAV DETR model has a training loss of 0.11. The significant decrease in training loss for our model shows its high efficiency and efficacy in acquiring features from the data. The results highlight the importance of using RGBIR-fused optical flow data to improve the overall performance of the proposed model compared to other methods.

Moreover, the experiment rigorously examined the entire processing time for the fusion method, Quickshift algorithm, and Graph Convolutional Network (GCN). More precisely, the fusion technique required a processing time of 17 milliseconds, the Quickshift algorithm cost 130 milliseconds, and the GCN algorithm utilized roughly 103 milliseconds. The total processing time for all three parts is 250 milliseconds. This detailed timing study provides valuable insights into the computational efficiency of each part, facilitating the evaluation of their respective contributions to the overall time required for processing.

The proposed OF-GPRN relies on optical flow-assisted graph-pooling and residual networks to enhance UAV detection. One limitation is its sensitivity to changes in environmental conditions, such as abrupt rain, which creates noises in the image backgrounds, and it would be challenging for

the optical flow to identify the UAVs. In our future work, we could explore strategies to improve the algorithm's robustness under varying conditions.

Moreover, the effectiveness of the OF-GPRN system is demonstrated using a benchmark UAV catch dataset. However, the algorithm's performance may be influenced by variations in different UAV shapes, tiny sizes, and very fast motion patterns not covered comprehensively in the training data. Discussing strategies for handling diverse UAV scenarios and potential limitations related to data variations would strengthen the work. The proposed system is evaluated on a specific dataset, and its generalization to new data environments may be a concern due to the feature variations. The OF-GPRN system model uses a multi-stage approach that includes fusion of RGB and IR signals, optical flow processing, and the Graph-Pooling Residual Network. While these stages contribute to improved detection rates using high-computational real-time ground-based static servers, the OF-GPRN system model may face computational complexities, especially in real-time applications with onboard devices. Addressing the trade-off between accuracy and computational efficiency could be an important consideration.

4.3 Chapter Summary and Concluding Remarks

In this chapter, we propose a new OF-GPRN system to enable the precise detection of UAVs in dual day and night visions that suffer from time-varying lighting conditions and high background similarities. The proposed OF-GPRN system incorporates optical fusion techniques and effectively eliminates extraneous backgrounds, resulting in enhanced clarity of RGB/IR imaging. Moreover, the GRSaN is extended in the OF-GPRN system to stabilize the learning capability, improve feature learning during training, and reshape the UAV.

In addition, the OF-GPRN system extracts the pixels-to-node representation of the adjacency matrix to achieve an accurate graph with information about the pixels of the conspicuous frame and has the precise learning ability to capture correlations among the pixels in the fused images to identify the suspicious UAV. Experimental results show that our proposed OF-GPRN system achieves an impressively low loss of 0.026. Compared to previous ResGCN object detectors, which recorded an mAP of 69.9%, the OF-GPRN system delivers superior performance, reaching an mAP of 87.8% on the demanding RGB-IR-based benchmark UAV catch dataset.

Moreover, optimization of the model should be the objective of future studies for onboard real-time system models like UAVs. In addition, optical fusion limits the present study to static cameras; we aim to improve the model for implementation with moving objects, such as autonomous vehicles and UAVs. Research on its possible applications in different environmental circumstances, such as rain and different types of UAVs, is required. Improving accuracy and efficiency requires testing on more and more diverse datasets.

Chapter 5

Latency Efficient Edge Vision Based System for UAV2UAV Detection

In this chapter we explore **Challenge 3-** with novel Graph Convolutional Network-Assisted Deep Reinforcement Learning model (GCN-Assisted-A2C) to manage the sizes of the sub-area video frames, which impacts the packet size and creates transmission delay, as shown in Fig. 5.1. The key of the system model is GCN, which is capable of selecting the region-of-interest area of small objects and its neighboring group of pixel regions from a video frame that have higher correlations with the centroid of the region-of-interest. However, it is important to select a group of pixels-correlated regions from the output of the GCN module to optimize the latency unless the accuracy of object identification is satisfied.

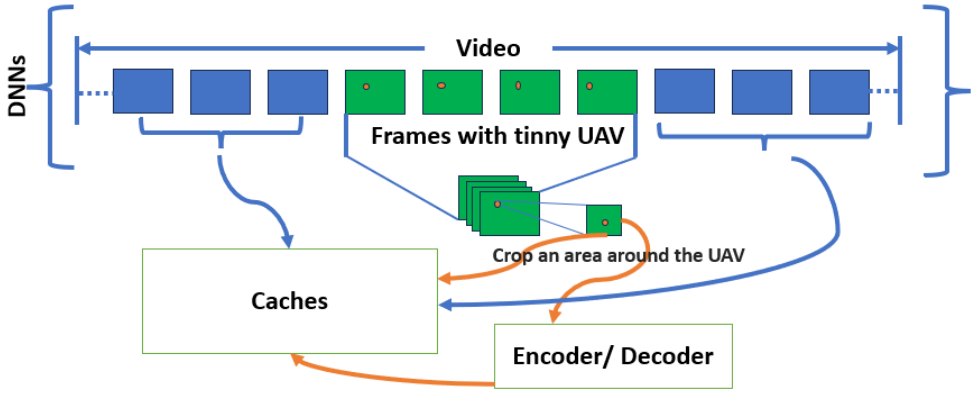


Figure 5.1: On-board UAV processing video to detect suspicious UAVs.

For this, we propose a novel GCN-Assisted-A2C that is fast, accurate, and flexible with the dynamic movement of the UAV on board and capable of selecting the region-of-interest area of small objects and its neighboring group of pixel regions from a video frame that has higher correlations with the centroid of the region-of-interest. Our objective is to efficiently optimize and control the

transmission delay in order to accurately identify UAVs. The YOLO model in the backend defines the bounding box for the UAV, and GCN takes advantage of the bounding box center to find the group of pixels-correlated regions with a strong relationship with the center. While the A2C module of the proposed GCN-Assisted A2C model selects the group of pixels-correlated regions found by GCN to send it to the server. The server encoder finds the object and sends back the information of accuracy and latency to the GCN-Assisted-A2C model to optimize the latency efficiently in reducing the transmission delay to select a minimum group of pixels-correlated regions.

In summary, the main contributions of this paper are the following:

- We propose a novel GCN-Assisted-A2C model that takes the input of the detection YOLO algorithm, uses the centroid of the detection object area, and creates clusters using the GCN of feature-correlated groups of pixels and regions of hidden pixel relationships. The GCN-Assisted-A2C model utilizes a SLIC segmentation of frames as ground truth for the groups of pixel regions during training the GCN model.
- The proposed GCN-Assisted-A2C model develops A2C that uses the GCN as a state based on a group of pixels correlation values of future predictions to process the output to take action and transmit it to the server. The A2C reward is also based on the Lagrangian dual form for checking the sensitivity of the objective function. The aim is to reduce the risk of temporal difference learning error estimation for both the critic and the actor loss to update the policy. While we used dual gradient descent to update the λ variable parameter and avoid lack of latency convergence and over and under penalization constraint violation during latency optimization.
- We evaluated the proposed GCN-Assisted-A2C model on the YouTube dataset to assess its real-world performance and compared it with three benchmark models: FlexPatch, Edge-Duet, and DRL. The proposed model achieved an optimal inference latency of 45 ms and an average precision of 70.72%, with a mean IoU of 60.30%, indicating more accurate RoI region selection and higher overall detection performance compared to the benchmark methods.

Beyond the proposed GCN-assisted A2C, this work uses the following technologies in the end-to-end video pipeline:

- **UAV side:** An onboard camera combined with YOLO-Nano inference for real-time object detection. When the detection confidence falls below a predefined threshold, region-of-interest (RoI) image crops are extracted and transmitted to the edge server using an intra-coded image encoding scheme (H.264 Intra-only), enabling efficient and robust transmission of independent RoI frames.
- **Edge server side:** ESRGAN to regenerate missing contextual information, FFDNet for image denoising, Zero-DCE for low-light enhancement, followed by a large-scale YOLO model for accurate object detection and refinement.

5.1 Mathematical System Modeling

We consider a sequence of encoded video frames $\{F_t\}$ of spatial size $W \times H$ pixels captured by a UAV. A light onboard detector (e.g., YOLO-Nano) produces at times t a bounding box $b_t = (x_c, y_c, \zeta_W, \zeta_H)$ with center (x_c, y_c) and size (ζ_W, ζ_H). ζ_W and ζ_H are the width and height of the detected bounding box. The transmitter chooses an action E_t that parameterizes a region of interest $f_t(E_t)$ centered at (x_c, y_c) with scale factors α_t and β_t (Eq. 5.1).

$$f_t(E_t) = \left[x_c - \frac{\alpha_t \zeta_W}{2}, x_c + \frac{\alpha_t \zeta_W}{2}\right] \times \left[y_c - \frac{\beta_t \zeta_H}{2}, y_c + \frac{\beta_t \zeta_H}{2}\right] \quad (5.1)$$

Let $|f_t| = \alpha_t \beta_t \zeta_W \zeta_H$ be the RoI area in pixels. For symmetric expansion, we consider $\alpha_t = \beta_t = 1 + 2E_t$, where E_t is the scaling factor that determines the cropping region of the bounding box size, with $E_{\min} \leq E_{t,n} \leq E_{\max}$.

In practice, each frame F_t will contain several sub-areas of interest (i.e., the RoIs). Let $\mathscr{F}_t = \{f_{t,1}, \ldots, f_{t,n}\}$ be the set of such sub-areas in which $f_{t,n}$ denotes the n-th sub-area of frame F_t and N the set size. Once the N RoIs are defined, they are transmitted through the communication channel. Of particular importance to the transmission time is the RoIs area, which is given by $|f_{t,n}| = (1 + 2E_{t,n})^2 \zeta_W \zeta_H$, because it determines the amount of information to be transmitted. Without loss of generality, we consider that each RoI is transmitted inside one communication packet of size $S_{t,n}$ as given by Eq. 5.2, where ρ (bits/pixel) captures the codec rate at the chosen quality and Ψ_{frame} aggregates headers and per-packet overhead.[1]

$$S_{t,n} = \underbrace{\rho\,|f_{t,n}|}_{\text{content payload}} + \underbrace{\Psi_{\text{frame}}}_{\text{headers/overhead}} \quad (5.2)$$

Eq. 5.2 shows the case of successful packet transmission. However, in real-time the communication between the UAV and the Edge server is affected by packet loss, which in turn triggers retransmissions, impacting the RoIs transmission times. This loss-induced extension of the transmission times must be considered in the target delay constraint (Wu et al., 2019).

Therefore, we define the expected transmission time per packet $T_{t,n}$ as given by Eq. 5.3 which is represented for transmission in (Yu et al., 2022; Wu et al., 2017), where B is the size of the negative acknowledgment packet $(NACK)$, R (bits/s) is the wireless link transmission rate, $\min RTT$ is the round-trip time, and $q_{t,n}$ is the per-packet success probability (after PHY/MAC). We also consider a stop-and-wait ARQ retransmissions protocol with a NACK packet of size B bits.

[1] If tiling is used, Ψ_{frame} can include tile headers; ρ can vary mildly with texture/motion.

$$T_{t,n} = \underbrace{\left(\frac{S_{t,n}}{R} + \frac{\min RTT}{2}\right)}_{\text{first attempt}} + \underbrace{\frac{(1-q_{t,n})}{q_{t,n}}}_{\text{expected retries}} \cdot \underbrace{\left(\frac{B}{R} + \min RTT + \frac{S_{t,n}}{R}\right)}_{\text{each retry}}, \tag{5.3}$$

5.1.1 Accuracy Constraints Strategy

We now construct the constraints for our framework regarding detection accuracy and sub-area frame size. Emphasizing the importance of carefully selecting appropriate frame cropping area scales for UAV detection accuracy greatly influences performance. The necessity for careful sub-area frame scale selection arises from two primary factors: first, real-world UAVs come in a wide range of sizes, often unfamiliar to the visual system; and second, the distances between objects and the camera may vary, potentially lacking prior information. We aim to achieve higher accuracy $(\mathscr{A}_s)$ on the server side when receiving sub-area frames $f_{t,n}$. The accuracy $\mathscr{A}_s$ is defined as a function of ROI size, as shown in Eq. 5.4 below, with the following condition.

$$\mathscr{A}_s^{(k_0)} \;=\; \mathscr{P}(|f_{t,n_0}|) \;=\; \mathscr{P}\Big((1+2E_{t,n_0})^2\,\zeta_W\,\zeta_H\Big). \tag{5.4}$$

$$\mathscr{A}_s^{(k+1)} = \begin{cases} \mathscr{P}(|f_{t,n+1}|), & \text{if } \mathscr{A}_s^{(n)} \le A_{threshold}, \\ \mathscr{A}_s^{(n)}, & \text{otherwise.} \end{cases} \tag{5.5}$$

Eq. 5.5 states that if $\mathscr{A}_s$ is less than or equal to the threshold $A_{threshold}$, then the server-based UAV detection accuracy $\mathscr{A}_s$ can be increased by choosing an optimal scaling factor $E_{t,n}$. While $\mathscr{P}\{\cdot\}$ is a probability of each transmitted subarea, which is $\mathscr{P}\{\text{correct detection} \mid f_{t,n}\}$ how much the likelihood that the server correctly detects the UAV given the received RoI. Here, the constraint $\mathscr{A}_s \ge A_{\text{limit}}$ must be satisfied to achieve the desired accuracy. To achieve this, the scaling factor is updated as $E_{t,n+1} = V_E \cdot E_{t,n}$, where V_E denotes the variation in size between sub-areas $f_{t,n}$ and $f_{t,n-1}$. The new scaling factor $E_{t,n+1}$ is applied to define the next sub-area $f_{t,n+1}$, which has a larger area than intial $f_{t,n}$ of n_0. The variation V_E between two consecutive sub-areas is computed as in Eq. 5.6. This equation computes the absolute difference between the change in scaling factors $(E_{t,n-1} - E_{t,n})$ and the product of ζ_W and ζ_H.

$$V_E \;=\; \left|(E_{t,n-1} - E_{t,n}) \cdot (\zeta_W \cdot \zeta_H)\right|, \tag{5.6}$$

This absolute difference quantifies the extent to which the value of V_E is altered as a result of variations in $E_{t,n}$, while simultaneously considering the dimensions of the cropped sub-area of the frame. The absolute value function ensures that the result is always non-negative.

Moreover, $\mathscr{P}\{\cdot\}$ must be optimal to obtain maximum accuracy and low latency. A retrieval process based on $E_{t,n}$, treated as a ranking or sorting method for different sub-area frames, does not

work because it is non-differentiable. In addition, linear optimization cannot be applied, since constant changes in the input distance $|f_{t,n}|$ cause the $\mathscr{A}_s$ values to be discontinuous (He et al., 2018). In order to obtain an optimal $\mathscr{P}\{\cdot\}$, groups of pixel correlations between regions must be considered, since correlations between center and neighboring pixels are essential. This approach also reduces latency by enabling a one-time prediction of a sub-area of the frame rather than relying on a sequence of actions. The GNN primarily focuses on prediction and determining the optimal decision for a single occurrence. Using the GNN algorithm to identify the area according to pixel correlations results in a more significant decrease in latency compared to linear optimization. This is because linear optimization is based only on $E_{t,n}$, which does not guarantee higher precision in the first transmission of sub-area frames and is vulnerable to selecting inappropriate regions of the frame with low pixel correlations.

5.1.2 Problem Formulation

The above analysis shows that the number of transmissions of sub-area frames $f_{t,n}$ in $T_{t,n}$ is critical to maintaining the delay below the desired threshold of a single full-frame transmission delay. According to Alhilal et al. (2022), the total transmission time of frames can be expressed as the sum of all single sub-area frame delays, which is given by Eq. 5.7.

$$\sum_{n=1}^{N} T_{t,n} = \sum_{n=1}^{N} \left(\frac{S_{t,n}}{R} + \frac{\min RTT}{2} + \frac{(1-q_{t,n})}{q_{t,n}} \left(\frac{B}{R} + \min RTT + \frac{S_{t,n}}{R} \right) \right) \tag{5.7}$$

The optimization problem to minimize the sum of the total delay can be stated as follows. Given the estimated packet sizes $S_{t,n}$ for each transmission at send rate R and the one-way propagation/processing delay $\frac{\min RTT}{2}$, the objective is to achieve minimum transmission latency by adapting each cropped sub-area $f_{t,n}$ via a variable scaling factor $E_{t,n}$ per packet. The design also accounts for the number of sub-area transmissions and leverages an optimal $\mathscr{P}(.)$ (pixel-correlation-based precision probability), so that the server-side accuracy $\mathscr{A}_s$ for each received sub-area frame is high and stable. This is formally expressed in Eq. 5.8,

$$\textbf{OP}: \quad \arg \min_{\{f_{t,n}\}_{n=1}^{N}} \sum_{n=1}^{N} T_{t,n} \tag{5.8}$$

$$\text{subject to:} \quad \mathscr{A}_s \geq A_{\text{threshold}}, \quad n = 1, \ldots, N.$$

where

$$
\begin{aligned}
T_{t,n} &= \left(\frac{S_{t,n}}{R} + \frac{\min RTT}{2}\right) + \frac{1-q_{t,n}}{q_{t,n}} \\
&\quad \left(\frac{B}{R} + \min RTT + \frac{S_{t,n}}{R}\right), \\
S_{t,n} &= \rho\,|f_{t,n}| + \Psi_{\text{frame}}, \\
\mathscr{A}_s &= \mathscr{P}(|f_{t,n}|), \\
E_{t,n+1} &= V_E \cdot E_{t,n-1}, \\
V_E &= |(E_{t,n-1} - E_{t,n}) \cdot (\zeta_W \cdot \zeta_H)|\,.
\end{aligned}
$$

5.1.3 Complexity

Problem **(OP)** induces a constrained Markov decision process (MDP) with continuous action variable $E_{t,n}$ and a non-convex, data-driven accuracy function $\mathscr{A}_s(\cdot)$. Exact dynamic programming is intractable. Therefore, we adopt a model-free primal–dual actor–critic method that learns policies minimizing the expected delay while satisfying accuracy constraints. In the simplified case where $E_{t,n}$ takes values on a discrete grid of K levels, and both $\mathscr{A}_s$ and $T_{t,n}$ are monotone in $E_{t,n}$, each sub-area frame can be optimized independently in $O(NK)$ by scanning the grid. However, when temporal correlations, joint constraints, and stochastic packet success are considered, the problem becomes a high-dimensional constrained MDP, making reinforcement learning methods more appropriate.

5.2 Proposed GCN-Assisted DRL System Model

This section presents the model of the GCN-assisted A2C system. The proposed model employs GCN to extract the areas of features that have a strong relationship with the object center, as well as the strong correlation between groups of pixels to facilitate future prediction. While A2C exploits the learning outcomes of GCN to train the action of latency optimization and accuracy constraint satisfaction.

In the previous section, we formulated the transmission optimization problem (OP) by defining the action as the ROI scaling factor $E_{t,n}$, which directly adjusts the size of the cropped sub-area frame $f_{t,n}$. While this abstraction provides a tractable CMDP formulation, it does not exploit correlations between pixels inside the bounding box, nor does it account for complex feature distributions that arise under different UAV sizes, lighting conditions, and motion. To address these limitations, we extend the baseline model with a GCN-assisted A2C framework. In this setting, the GCN predicts correlated pixel groups $\mathbb{S}_i$ that summarize feature relationships in the frame, and these groups are mapped into bounding boxes $\mathbb{B}_i$ centered around correlation-weighted centroids. Each $\mathbb{B}_i$ corresponds to an ROI that can be expressed as $f_{t,n}$ for some effective scaling factor $E_{t,n}$. In other words, the GCN refines the baseline $E_{t,n}$-based action by constraining the A2C policy to choose

feature-correlated regions that are both latency-efficient and accuracy-preserving. The following presents the proposed GCN–A2C solution model in detail.

5.2.1 CMDP Formulation and Information Flow

At each time step t, the Constrained Markov Decision Process (CMDP) state is defined as $s_t = (\phi_t^{\text{det}}, \mathbf{g}_t, q_t, minRTT_t, \lambda_t)$, where ϕ_t^{det} are detector features (bounding boxes, scores), $\mathbf{g}_t$ is a GCN embedding summarizing pixel-group correlations around (x_c, y_c), and λ_t is the CMDP multiplier for latency–accuracy constraints.[2]

In the proposed system model, the feature vector $(\mathbb{V})$ for GCN obtained from pixel-group sets $\mathbb{S} = \{\mathbb{S}_i\}_{i=1}^{Z}$ to form the GCN embedding $\mathbf{g}_t$. That is, $\mathbf{g}_t = \mathbb{V}$ provides a compact encoding of correlations between disjoint pixel groups, which the A2C agent uses to guide decision-making.

The CMDP action corresponds to selecting a bounding box $\mathbb{B}_i$ from the GCN-predicted correlated pixel group $\mathbb{S}_i^*$. Each bounding box $\mathbb{B}_i$ defines a sub-area ROI in the original OP formulation: $\mathbb{B}_i \equiv f_{t,n}$. Thus, the GCN transforms pixel correlations into bounding-box actions that the A2C policy transmits to the server, balancing latency $T_{t,n}$ and accuracy $\mathscr{A}_s$. After transmission, the server returns the observed accuracy (AP or IoU proxy) and measured delay, which update the estimated accuracy $\mathscr{A}_s$ and success probability q_t online.

5.2.2 GCN-A2C for Decision Making

We use an A2C strategy to optimize latency for solving the OP problem formulated from Eqs. 5.2 to 5.7 (Li et al., 2022c). Figure 5.2 illustrates a framework with three primary modules. The first is the **actor module**, which takes an action $\{\mathbb{B}_i, \mathbb{S}_i\}$. Here, the GCN predicts correlated pixel groups $\mathbb{S}_i$, and the corresponding bounding box $\mathbb{B}_i$ defines the cropped sub-area. Formally, each $\mathbb{B}_i$ corresponds to an ROI $f_{t,n}$ with scaling factor $E_{t,n}$, i.e., $\mathbb{B}_i \equiv f_{t,n}$. Without GCN, the A2C actor would need to search exhaustively over pixel groups, leading to high computation. The pixel-group shape adapts to the variance or distribution around the centroid and has width ζ_W and height ζ_H. The **critic module** evaluates $\{T_{t,n}, \mathscr{A}_s\}$, i.e., the transmission delay per sub-area packet and the accuracy (mAP proxy) achieved at the server. The **policy** is improved by combining A2C training with GCN predictions: the GCN records correlations in pixel groups and provides future predictions, which periodically update the actor to select feature-correlated regions that minimize latency while ensuring $\mathscr{A}_s$ constraints are met. This guarantees that the policy $\pi(a_t|s_t)$ adapts to varying UAV sizes and pixel correlations. The size of each pixel group is determined by $\Theta = \sqrt{\frac{K}{M_i}}$, where K is the total number of correlated pixels near the object center, and $M_i = |\mathbb{S}_i|$ is the cardinality of the subset used by the GCN to guide actor decisions. Exhaustively evaluating groups of size M_i requires $\binom{K}{M_i} = \frac{K!}{M_i!(K-M_i)!} = O(K^{M_i})$, making brute-force search infeasible. Instead, the GCN learns to predict future feature-correlated actions for

[2]Not all state variables appear directly in the OP formulation, which is expressed only in terms of the optimization variables $f_{t,n}$, $E_{t,n}$, $T_{t,n}$, and $\mathscr{A}_s$. Instead, the CMDP state aggregates the contextual information needed by the RL policy to generate decisions $f_{t,n}$ that solve OP in practice.

A2C, reducing complexity while satisfying $\mathscr{A}_s$. The proposed system not only selects pixel groups but also captures stronger correlations between the center and neighboring hidden pixels. In the next section, we detail the structure of the GCN model.

5.2.3 Hidden Pixels Relation Exploration

This section presents a detailed graphic representation along with A2C of the proposed graph neural network structure for higher accuracy and low latency, shown in Fig. 5.2. To find the high correlations between the pixels of an image for optimal sub-area frame selection, we need to divide the frame into different partitions. The frame partitioning process is a critical and fundamental stage within any low-level vision system to find the region of interest where pixels have higher correlations with each other. Segmentation refers to the procedure of dividing a frame into distinct areas that do not overlap, with the objective of ensuring that each zone shows correlations and that the separation between areas is abrupt. Different methods for pixel correlations have been published in the past. However, it remains a challenge to identify a specific approach that shows consistency throughout numerous lighting conditions or low-visibility conditions. In addition, CNN performs convolution operations in smaller, more uniform areas and obtains features at the pixel level. However, GCN performs convolution operations in large areas with unevenly correlated pixels and gets features at the aggregated pixel level. The GCN method uses the average of all pixels in a set to give strong edge weights to areas with similar correlations and weak or no edge weights to areas with no correlations (Li et al., 2023, 2022c).

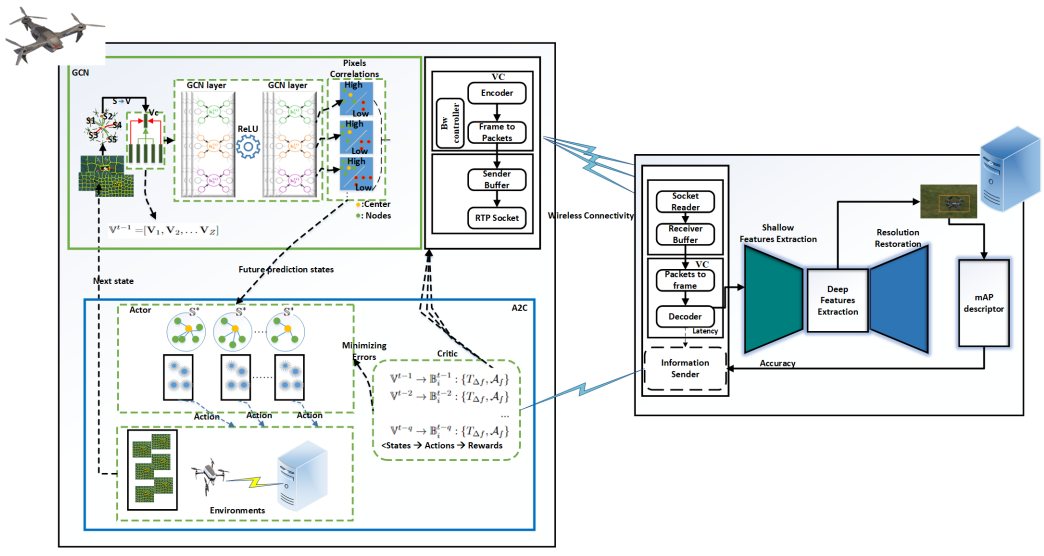


Figure 5.2: The figure illustrates the proposed framework, indicated by several color-coded parts. The left-side black box encloses the proposed model region, which symbolizes the core framework of the GCN-assisted A2C. The blue box shows the actor who selected the action. While the critic's role in action-value evaluations. The green box is the GCN future state prediction, which regulates actor behavior to optimize latency and minimize errors.

5.2.3.1 State

Let $\mathbb{S} = \{\mathbb{S}_i\}_{i=1}^{Z}$ represent the set of aggregate pixel groups forming the state, where $\mathbb{S}_i = \{\mathbf{Y}_i^j\}_{j=1}^{M_i}$, $M_i = |\mathbb{S}_i|$. Here, $\mathbb{S}_i$ denotes the i-th group of correlated pixels and M_i is its cardinality. The pixel groups are disjoint, i.e., $\mathbb{S}_i \cap \mathbb{S}_j = \emptyset$ for all $i \neq j$. Moreover, the total number of pixels is preserved: $\zeta_H \times \zeta_W = \sum_{i=1}^{Z} M_i$. Each group is summarized by its mean feature vector, yielding the state representation, which is given in (5.9):

$$\begin{aligned} \mathbb{V} &= [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_Z]^T \\ &= \left[\frac{1}{M_1} \sum_{j=1}^{M_1} \mathbf{Y}_1^j, \ldots, \frac{1}{M_Z} \sum_{j=1}^{M_Z} \mathbf{Y}_Z^j \right]^T, \end{aligned} \tag{5.9}$$

where $\mathbf{Y}_i^j$ is the j-th pixel feature in group $\mathbb{S}_i$. Thus, $\mathbb{V}$ is a column vector of dimension Z, with each entry being the average feature of a pixel group. This representation captures the correlation structure of the aggregated pixel regions, which serve as the input nodes for the GCN. After processing the initial state, the GCN predicts future correlation patterns, which guide action selection by cropping the corresponding bounding box region. A focal classification loss $\mathfrak{L}_{cls}$ is used to train the GCN for UAV pixel correlation prediction, which is given in (5.10):

$$\mathfrak{L}_{cls} = \lambda_1 \mathfrak{L}_{tcrp} + \lambda_2 \mathfrak{L}_{tcrn}, \tag{5.10}$$

where $\mathfrak{L}_{tcrp}$ measures correlation consistency among pixels inside the UAV bounding box, while $\mathfrak{L}_{tcrn}$ penalizes spurious correlations in background regions outside the UAV. The loss $\mathfrak{L}_{tcrn}$ suppresses background noise near the UAV centroid, allowing the GCN to emphasize UAV-related pixels. Here, λ_1 and λ_2 are tunable weighting coefficients used to balance the influence of $\mathfrak{L}_{tcrp}$ and $\mathfrak{L}_{tcrn}$ during training.

5.2.4 A2C for T_f Optimization with $\mathscr{A}_s$ Satisfaction

5.2.4.1 Action

In the previous section, we proposed a GCN method to capture correlations between pixels by partitioning images into groups of pixels. Using this foundation, we now leverage the correlation data to define the **action** in the A2C framework, which specifies the region to be transmitted. Let the action be a bounding box $\mathbb{B}_i$ associated with a correlated pixel group $\mathbb{S}_i^*$ predicted by the GCN. The goal is for the GCN to identify highly correlated pixel groups that can be mapped into compact rectangular or square regions. To achieve this, we define a correlation-weighted centroid (x_c, y_c) based on the coordinates (x_Φ, y_Φ) of pixels $\Phi \in \mathbb{S}_i^*$ with weights $\mathfrak{C}_\Phi$ (Eq. 5.11). $\mathfrak{C}_s$ denotes the correlation weight of pixel s, highlighting its contribution to the centroid.

$$x_c = \frac{\sum_{\Phi \in \mathbb{S}_i^*} \mathfrak{C}_\Phi \cdot x_\Phi}{\sum_{\Phi \in \mathbb{S}_i^*} \mathfrak{C}_\Phi}, \quad y_c = \frac{\sum_{\Phi \in \mathbb{S}_i^*} \mathfrak{C}_\Phi \cdot y_\Phi}{\sum_{\Phi \in \mathbb{S}_i^*} \mathfrak{C}_\Phi}. \tag{5.11}$$

Using the centroid, the bounding box action $\mathbb{B}_i$ is defined as in Eq. 5.12, where ${\zeta^*}_W$ and ${\zeta^*}_H$ denote the width and height of the correlated region.

$$\mathbb{B}_i : \left[x_c - \frac{{\zeta^*}_W}{2}, y_c - \frac{{\zeta^*}_H}{2}, x_c + \frac{{\zeta^*}_W}{2}, y_c + \frac{{\zeta^*}_H}{2} \right], \tag{5.12}$$

These dimensions are adaptive, depending on the variability of the pixel-group distribution around the centroid. This flexibility improves the expressiveness of the bounding box and ensures that correlated regions are captured accurately.

5.2.4.2 Reward

The A2C training's reward (r_t^{A2C}) determined by the totals of sub-area frames transferred to the server within the action on the group of pixel regions as $\left\{ \mathbb{S}^{\subsetneq}{}_i \mid \mathbb{S}^*{}_i, \mathbb{B}_i \right\}$. As $\mathbb{S}^*{}_i$ represents the correlated group of GCN-trained pixels. The action $\mathbb{B}_i$ is taken by the UAV, and the GCN-predicted group of pixels transits from $\mathbb{S}^*{}_i$ to $\mathbb{S}^{\subsetneq}{}_i$. However, the reward can only be optimized when $\mathscr{A}_s \geq A_{threshold}$ accuracy is considered for positive detection of the object on the server. It is essential to achieve optimal transmission latency and avoid penalization for constraint violation. It might be possible to put a condition for $\left\{ \mathbb{S}^{\subsetneq}{}_i \mid \mathbb{S}^*{}_i, \mathbb{B}_i \right\}$ preserving the feature-correlated region that can be achieved by Lagrangian dual form, as given by Eq. 5.13 where λ shows how sensitive the objective function is to changes in the value of $\mathscr{A}_s \geq A_{\text{threshold}}$, the accuracy constraints.

$$\mathscr{L}(\mathbb{S}^{\subsetneq}{}_i, \mathbb{S}^*_i, \mathbb{B}_i, \lambda) = \min_{\forall i_{\mathbb{S}^*{}_i} \in [1,Z]} T(f_{t,i_{\mathbb{S}^*_i}}) + \lambda (A_{\text{threshold}} - \mathscr{A}_s(f_{t,i_{\mathbb{S}^*_i}})), \tag{5.13}$$

By using this approach, we are able to ensure that minimizing the objective function will only occur if the accuracy requirements are achieved. Otherwise, OP is penalized and stops optimizing more during $\mathscr{A}_s < A_{\text{threshold}}$. The fixed λ creates a lack of convergence and over- and under-penalization constraint violation during latency optimization. To address these issues, we use dual gradient descent to dynamically update the λ as in Eq. 5.14 where η is the learning rate, which is initially set to 0.01 with a weight λ of 1.

$$\lambda \leftarrow \lambda + \eta (A_{\text{threshold}} - \mathscr{A}_s), \tag{5.14}$$

The η parameter dynamically adjusts λ during constraint violation. So $r_t^{\text{A2C}} \left\{ \mathbb{S}^{\subsetneq}{}_i \mid \mathbb{S}^*{}_i, \mathbb{B}_i \right\}$ of the training of A2C is determined by offloaded group of pixel sets to the server and is given by Eq. 5.15.

$$r_t^{\text{A2C}} \left\{ \mathbb{S}^{\subsetneq}{}_i \mid \mathbb{S}^*{}_i, \mathbb{B}_i \right\} = \min_{\forall i_{\mathbb{S}^*{}_i} \in [1,Z]} T(f_{t,i_{\mathbb{S}^*_i}}(\mathbb{S}^*_i)). \tag{5.15}$$

We incorporate the reward into the temporal difference learning (δ_t) error for the critic and actor loss as in Eq. 5.16 where $V^*(\Phi_{t+1})$ is the update estimated state of Φ_t that is the old estimated state of $V^*(\Phi_t)$.

$$\delta_t = r_t^{\text{A2C}} \left\{ \mathbb{S}^{\subsetneq}{}_i \mid \mathbb{S}^*{}_i, \mathbb{B}_i \right\} + \gamma V^*(\Phi_{t+1}) - V^*(\Phi_t), \tag{5.16}$$

The critic loss is evaluated by squaring the (δ_t) while the actor loss is based on the advantage function to regulate the policy $(\pi_\theta(a_t \mid \Phi_t))$ and map the states (Φ_t) to the actions (a_t) as given by $-\log \pi_\theta(a_t \mid \Phi_t) \cdot \delta_t$. The part $-\log \pi_\theta(a_t \mid \Phi_t)$ takes action under the current policy π_θ for the state Φ_t and takes the product of the likelihood of action with δ_t to update the policy using Eq. 5.17, where $\mathscr{W}_c$ shows the weighted factor for the critic loss.

$$\mathscr{L}_{loss} = -\log \pi_\theta(a_t \mid \Phi_t) \cdot \delta_t + \mathscr{W}_c \cdot (\delta_t)^2, \tag{5.17}$$

$\mathscr{L}_{loss}$ defines that the actor is maximizing the advantage and critic is minimizing the δ_t error with current updates within the constraints defined by A2C reward $r_t^{\text{A2C}} \left\{ \mathbb{S}^{\subsetneq}{}_i \mid \mathbb{S}^*{}_i, \mathbb{B}_i \right\}$.

5.2.4.3 Policy

The policy $\pi_\theta(a_t|\Phi_t)$ represents the A2C actor network that maps the CMDP state Φ_t to a probability distribution over actions a_t. Each action corresponds to selecting a bounding box $\mathbb{B}_i$, which defines the ROI $f_{t,n}$ to be transmitted. The policy parameters θ are updated using the actor loss in Eq. 5.17, which leverages the advantage signal δ_t to encourage actions that reduce latency $T_{t,n}$ while maintaining the required accuracy $\mathscr{A}_s$.

5.3 Experimental Work

5.3.1 Implementation of GCN-assisted A2C Model

This section presents an overview of the results for detection onboard the UAVs through the proposed GCN-assisted A2C model. The proposed GCN-assisted A2C model is built on a 64-bit Ubuntu 22.04 Linux workstation using Python 3.9 and PyTorch with a GPU of A300. In this study, we used YouTube videos for UAV-to-UAV monitoring. There are a total of 100 clips, all of which are 240 to 1080 resolution videos. These videos feature a variety of backgrounds, such as clouds, buildings, mountains, and sea, as well as fast movement, out-of-focus, and a range of sizes. The UAVs seen in each video come in a range of sizes, from large to small. We selected UAVs with a size range of 300mm to 1200mm to train and validate our proposed model. Furthermore, the UAVs cruise at speeds between 50 and 100 miles per hour, occasionally stopping altogether. We used all the validation videos from the YouTube stream for model training. We train the model

on all 100 videos but 70 of the videos, using the remaining 15 for testing and the remaining 15 for validation.

The GCN-assisted A2C is trained for 1 iteration per epoch to find the correlation between the different regions in every frame, and the GCN-assisted A2C is trained for 100 epochs to extract the pixel feature-correlated regions to send it to the server. GCN-assisted A2C is analyzed every 5 epochs using a validation function during the training process. This function finds the training loss and the number of iterations at which convergence occurs. The GCN-assisted A2C model with the lowest cost, denoted as the sub-optimum, is then used to assess the test data.

The GCN-assisted A2C used multiple layers that are organized into the GCN layers and multi-layer perceptron layers. The GCN layers module consists of three GCN multi-head layers, each containing two GCN layer edge softmax heads. The first GCN multi-head layer utilizes heads that apply a linear transformation f to convert input pixel features into output-correlated pixel features. Additionally, another linear transformation is used to convert input pixel features into output pixel feature-correlated features. The second layer undergoes a transformation where 128 input features are converted to 64 output pixel features, and 128 input pixel features are converted to 1 output pixel feature in a region that is correlated. The third layer of the GCN model performs a similar transformation by changing 256 input layer pixel features into 64 output pixel features and 256 input pixel features into 1 output pixel feature-correlated group of regions. The A2C module has two linear layers, in which the first layer maps 256 input pixel features to 128 output pixel features. For the final prediction, the second layer maps 128 input features to output feature-correlated group regions. It does this by using Eqs. 5.11 and 5.12 for the centroid.

The calculation of the policy gradients after multiplying by the benefits is known as policy loss. The Unmanned Aerial Vehicle (UAV) activities undergo training in the Advantage Actor-Critic (A2C) module for a total of 1500 steps, using a discount factor of 0.99. The loss function is calculated by taking the average of the squared differences between the estimated values and the actual returns. The loss calculation assigns a coefficient of 0.5 to the value function.

5.3.2 Ablation Study

In Fig. 5.3 we can see how well the proposed GCN-assisted A2C system model and different DRL algorithms work in terms of transmission latency rate during training. We found that the A2C, DDPG, and DQN models mostly do the same thing, which is to minimize transmission latency. However, the synchronous nature of the A2C model leads to higher performance than the deterministic nature of the DDPG and DQN models. The frame area's importance and the dynamical changes in the UAV's size due to different movements make A2C more stable than the DDPG and DQN models. In addition, cropping the frame area makes a clear action space, which might be one reason why A2C is generally more useful than DDPG. The frame area cropping creates priority fluctuation that leads A2C to better decision-making. It's hard for the DDPG and DQN models to make the best decision when there are a lot of different environmental factors

and frame area cropping variability problems. However, to adopt and find feature correlations and dependencies between different regions of the frame still reduces the performance of A2C. We proposed a GCN-assisted A2C system model, which ensures and preserves critical contextual features of groups of neighboring pixels in regions where A2C lacks the ability to deal with such dependencies. A2C processes and transmits a single sub-area frame entity and makes it difficult to find the nuanced interply between different sub-areas of frame parts or sizes of UAVs, variations, or environmental factors. Moreover, GCN assists the A2C to extract informative features of the UAV using aggregation and propagation of features across graph nodes to make better predictions. Similarly, the GCN-assested A2C system model reduces the high dimensionality focus to the specific interdependent area to efficiently handle energy constraints or bandwidth limitations. Without GCN, A2C struggles with scalibility, and it leads to high computation and might be complex architectures needed. In addition, GCN-assisted A2C is more robust to noises or missing pixel feature information to aggregate neighboring nodes feature information to improve predictions and reduce the transmission latency. If a portion of a frame is corrupt or unclear with environmental effects, then GCN leverages the hidden feature pixel information of the surrounding group of pixels, improves the transmission latency, and preserves the prediction performance constraints.

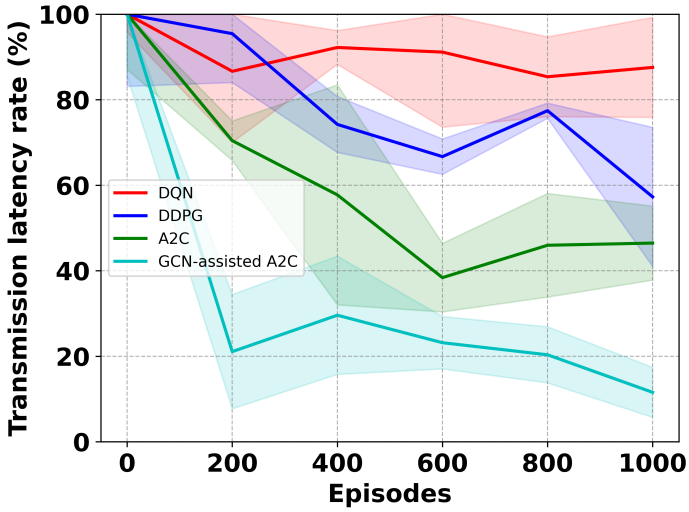


Figure 5.3: The ablation study of different RL algorithms performance using loss using data that have across different sizes of UAVs and dynamic movements with different environmental effects.

Moreover, in the ablation study, we used different DRL models (DQN, DDPG, and A2C) and studied required evaluation metric performance, like latency for each video frame transmission, overall AP, and mean IoU. The latency of the GCN-assisted A2C is much lower than that of the other benchmark DRL algorithms. The overall FPS transmission latency of the GCN-assisted A2C is 45 ms, which is much lower than the A2C without GCN, which has an FPS transmission latency of 170 ms. The GCN-assisted A2C achieved optimal latency due to the focus on the feature-correlated group of pixel regions to recommend to the A2C for optimal action. Moreover, Fig. 5.4

shows the AP and mean IoU of the GCN-assisted A2C and the three variants of the ablation study. The GCN-assisted A2C is the most effective across all the algorithms to transmit the optimal area of the frame that is utilized by the edge server. The GCN-assisted A2C achieved (AP @ 70.72%) and (mean IoU @ 60.30%) to improve (AP @ 62.80%) and (mean IoU @ 30.60%) over A2C. While A2C achieved (AP @ 43.44%) and (mean IoU @ 45.91%), which is higher in ratio than (AP @ 42.20%) and (mean IoU @ 60.50%) over the DDPG algorithm. Similarly, DDPG achieved (AP @ 30.55%) and (mean IoU @ 28.61%), and DQN with a very low score of (AP @ 17.30%) and (mean IoU @ 20.21%).

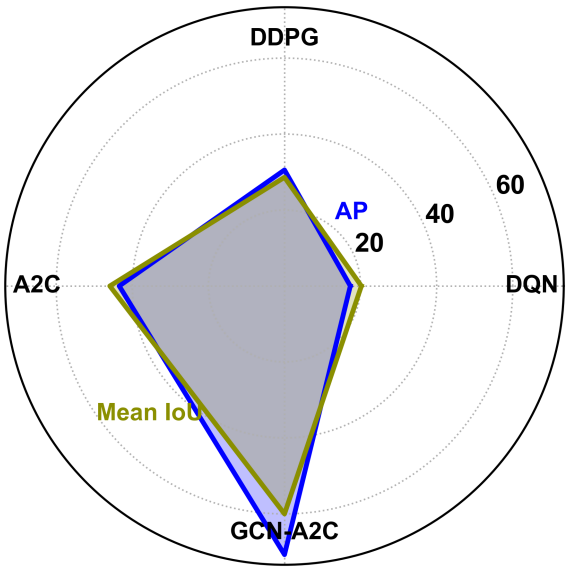


Figure 5.4: With the proposed GNN-A2C (GCN-assisted A2C) and different DRL algorithms, we can check the average (30FPS) IoU and tracking AP accuracy.

5.3.3 State-of-the-art (SOTA) comparisons

Fig. 5.5 shows the statistical performance of different state-of-the-art (SOTA) models, i.e., FlexPatch, EdgeDuet, and other DRL, to visualize the transmission latency in the range of cumulative distributed function (CDF) parameters. We found that the proposed GCN-assisted A2C had the best transmission latency of 50 ms with a standard deviation of 12 ms when compared to other SOTA models. FlexPatch came in second with a transmission latency of 90 ms and a standard deviation of 15 ms. Meanwhile, EdgeDuet achieved a video frame transmission latency of 110 ms, and its standard deviation is 20ms. In contrast, other DRL models achieve a video frame transmission latency of 177 ms, with a standard deviation of 33 ms. We can see from the CDF of the GCN-assisted A2C that the model works better for achieving the best video frame transmission latency across a range of communication factors.

At the same time, we examined how well the GCN-assisted A2C worked regarding transmission delay, AP, and meanIoU accuracy when compared to FlexPatch and EdgeDuet, using different

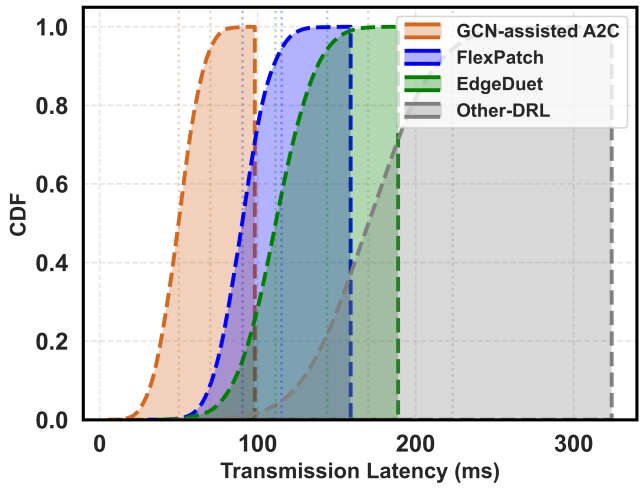


Figure 5.5: The effectiveness of GCN-assisted A2C over SOTA models to optimize the transmission latency.

video resolutions as shown in Fig. 5.6. In Fig. 5.6(a), we show how accurately we can find suspicious UAVs from start to finish, using AP and meanIoU accuracy. Similarly, in Fig. 5.6(b), the mean IoU over end-to-end latency is compared. The GCN-assisted A2C model outperforms previous models. For example, FlexPatch AP accuracy performance is 60.10% with meanIoU accuracy results of 48.42% and achieved minimum latency results of 90.10 ms. The FlexPatch adds overhead when scenes change rapidly or unpredictably, and it is highly challenging to select the optimal context-dependent features. EdgeDuet achieved AP accuracy of 42.35% with meanIoU accuracy results of 39.52% and achieved minimum latency results of 110.25ms. This model also uses many high-resolution tiles at the edge server and is unsuitable for real-time or low-bandwidth scenarios due to offloading frame tiles to the edge server. The GCN-assisted A2C model is better at handling quickly changing or unpredictable scenes because it uses groups of pixel areas as nodes and their connections to understand how they depend on each other. The GCN-assisted A2C model reached the best accuracy of 70.72% AP, with meanIoU accuracy results of 360.30%, and improved the transmission delay to 50ms. The GCN-assisted A2C model works well in real-time or low-bandwidth situations because it can handle parts of the frame by grouping together related pixel areas.

Fig. 5.6(c) shows the AP accuracy and meanIoU of the GCN-assisted A2C when transmitting videos of different frame sizes. The GCN-assisted A2C consistently achieves higher AP meanIoU accuracy at the server side. With the decrease in frame resolutions, GCN-assisted A2C accuracy drops upto 50 AP, which is our desired threshold to have a minimum of 50 AP at the server side.

5.3.4 Evaluation Results

We have qualitatively and quantitatively shown the expected trade-off between AP and latency improvement for three different UAV sizes under various circumstances. Each environment has three

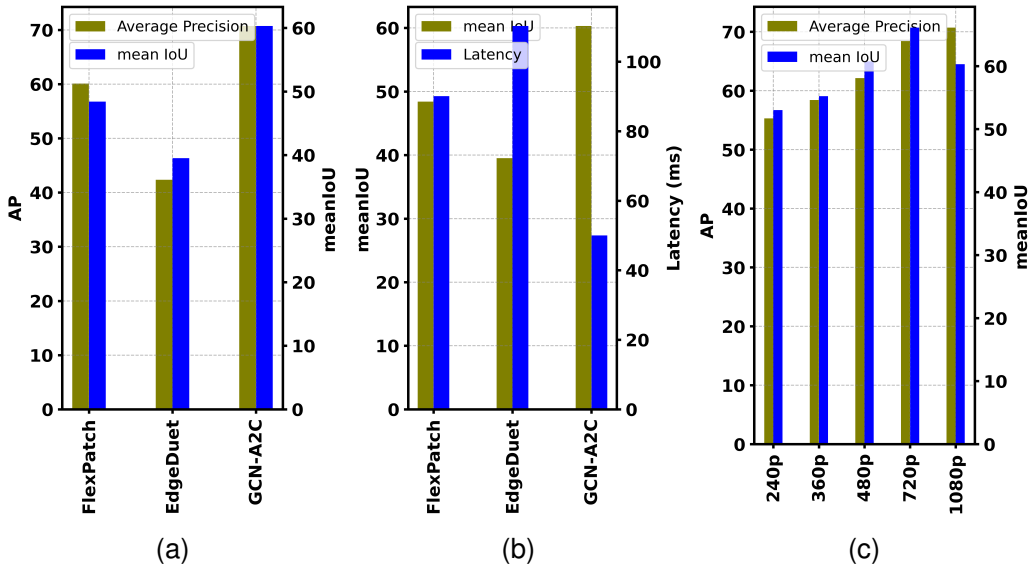


Figure 5.6: Performance of different models under UAV-to-server video sub-area of frame transmission. In (a), the AP and meanIoU accuracy of finding suspicious UAVs from end-to-end are shown. In (b), the meanIoU over end-to-end latency is compared. And in (c), the performance of GCN-assisted A2C is shown over different input video frame resolutions.

visual situations, with each subplot representing the AP and transmission latency. The blue bars show latency, whereas the olive-colored bars represent AP. Together, they exhibit both precision and efficiency. Overlaid visuals in each subfigure illustrate corresponding visual inputs related to the environmental condition. This shows how complex the agents' experiences are. Fig. 5.7 shows that smaller UAVs, similar backgrounds, and partial cloudiness can affect latency and reduce server-side AP. These unique features have a major effect on the efficiency of transmission in the other DRL models. The proposed GCN-assisted A2C keeps an AP above 50 while ensuring the best transmission speed in all conditions and sizes of UAVs during testing of the algorithms in three different conditions—cloudy, complex, and clear sky. This hypothetical situation motivates the advancement of smarter feature-correlated region prediction algorithms. The results indicate that GCN-assisted A2C consistently does better than the basic models in every situation, achieving a much higher AP with less delay. This benefit is especially evident in cloudy or messy situations, such as cloudy and complex situations, where classic models like DQN and DDPG do not perform well. GCN-assisted A2C, on the other hand, is robust to environmental changes and visual effects, making it more stable and suitable for a wider range of applications. In addition, GCN-assisted A2C continues to outperform its competitors in ideal situations such as clear skies. This demonstrates that it performs well in both challenging and normal circumstances. These findings show that GCN-assisted A2C could be effective for making real-time predictions based on vision when transmission speed and precision are important.

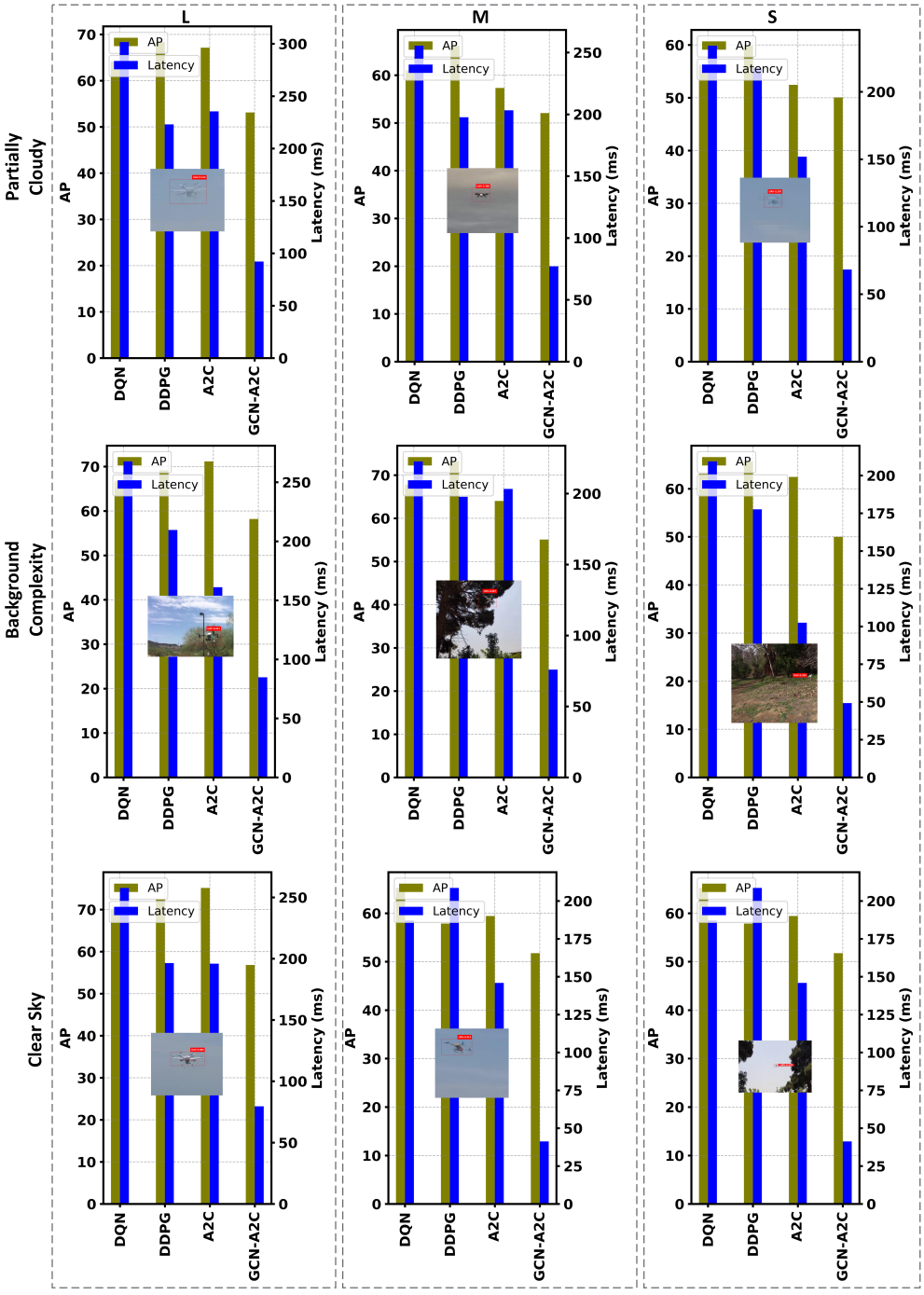


Figure 5.7: We evaluated three different sizes of UAVs—small (S), medium (M), and large (L)—under various conditions, such as clear sky, complex background, and partially cloudy.

5.4 Chapter Summary and Concluding Remarks

This chapter proposed a GCN-assisted A2C model that uses data from the UAV detection nano YOLO algorithm, focusing on the center of the detected object and the surrounding pixels. The design caters to scenarios in which on-board UAV detection accuracy is less than 50. We designed the detector model to identify suspicious UAVs in no-fly zones and pinpoint their location, utilizing the onboard UAV cameras. Next, we proposed using a GCN-assisted A2C to enhance the significance of certain pixel areas in the images by examining how each area connects to the center of the UAV, creating groups based on these connections, and calculating a total score with A2C. The nano detector model results emphasize that the GCN-assisted A2C model requires transmitting the frames to the server or not. We show that using the GCN-assisted A2C model can minimize the transmission latency and optimize until the AP is more than 50 on the server side. We conducted studies using authentic on-board UAV camera data and with different DRL models for the ablation study and SOTA models. Our GCN-assisted A2C model shows improved transmission latency and AP by using GCN to group related pixels and understand hidden pixel relationships, which helps the A2C make better decisions in different situations like clear skies, complex backgrounds, and partial clouds.

Chapter 6

Conclusion and Future Work

In last, we summarizes the dissertation results and findings and gives the prospective future direction to fulfill the limitations of proposed dissertation advancements. We highlights the contribution of our works and draw observations on it in Section 6.1. While delineates the future work and constraints of the proposed dissertation studies in Section 6.2.

6.1 Dissertation Conclusion

In this dissertation, we studied three different challenges. First, we investigated the UAV's detection in day and night dual vision in no-fly zones. Training a single deep learning model using multimodal inputs, such as RGB and infrared images, can introduce challenges due to differences in modality characteristics, which may lead to cross-modal feature inconsistencies during feature fusion. To mitigate this issue, we investigated the trade-off between RGB and IR features and employed a Bidirectional Feature Pyramid Network (BiFPN) to facilitate efficient multi-scale feature fusion and balanced feature propagation across modalities. We proposed a hybrid deep learning model that not only achieves the optimal accuracy for the correct localization of UAVs in day and night dual vision but also optimizes the training time and generalizes the model to different colors, positions of UAVs, and lighting conditions. Specifically, BiFPN combines the ideas from Feature Pyramid Network, Path Aggregation Network, and NAS-FPN in the form of multi-level feature fusion, which efficiently helps the proposed hybrid model extract features for both RGB and IR streaming. The photometric and geometric augmentation method, which includes techniques such as bounding box transformation, contrast, flapping, equalizing, solarizing, and sharpness, is used to enhance the efficiency and accuracy of the hybrid model under various background information and lighting conditions. To reduce the training time, we utilize cosine annealing. Cosine annealing gradually freezes the initial layers' weights, which preserves the RGB and IR features in the hybrid model. The experimental results show that the proposed hybrid deep learning model achieves 65% in terms of the mean average detection precision given the input videos in day and night vision.

Next, we investigated the effects of higher background noise and similarities on movable UAVs, which degrade the efficiency and accuracy of the deep learning model using RGB and IR static cameras. We studied how to fuse the IR and RGB images and subtract the background. We proposed OF-GPRN with optical fusion to remove superfluous backgrounds, which improves RGB/IR imaging clarity. Initially, a fusion module is developed in the proposed OF-GPRN to merge two input frames from RGB and IR sources while enhancing the quality of the output frames by reducing noise and adjusting illumination and color. We used optical flow BRAFT to monitor the UAV's movement and effectively approximate the actual physical motion projected onto the fused frame. The optical flow BRAFT provides concise information on the parts of the frame that are in motion and eliminates the background. Furthermore, OF-GPRN extends optical fusion by incorporating a graph residual split attention network and a feature pyramid, which refines the perception of UAVs, leading to a higher success rate in UAV detection. A comprehensive performance evaluation is conducted using a benchmark UAV catch dataset. The results indicate that the proposed OF-GPRN elevates the UAV mean average precision (mAP) detection rate to 87.8%, marking a 17.9% advancement compared to the residual graph neural network (ResGCN)-based approach.

Last but not least, people widely use UAV mobile onboard vision systems for monitoring various activities in no-fly zones and communicating with the ground server. We investigated that the latency of video transmission is one of the main factors that defines the speed of communication between the UAV and the server, and it needs to be optimized for low-latency communication between UAVs and the ground server. We proposed a GCN-assisted A2C system model to determine the optimal pixel-correlated area of a suspicious object. For the proposed system model, we employ the Lagrangian dual form with gradient descent to avoid convergence issues and violations of the over- and under-penalization constraint during latency optimization. Moreover, the proposed system model sends a subgroup pixel-correlated area of the frame from the UAV to the server rather than the transmission of the whole frame. The proposed framework utilizes the GCN model to explore hidden representations of feature-correlated groups of pixels. Moreover, GCN supervises the A2C model, which selects a subgroup to enhance transmission latency and oversees the training of UAV actions within the A2C framework. The experimental results indicate that GCN-assisted A2C achieves the lowest latency for video frame transmission and reduces the risk of false detection in UAV vision systems compared to other DRL and state-of-the-art models.

6.2 Future Work

In this section, we recognize the constraints of our work and propose strategies for its enhancement and expansion. We acknowledge these limitations and encourage further exploration of topics and research directions in the future. Next we are presenting the potential limitations and possible improvements in these research areas.

6.2.1 Dual Day Night Vision Systems

Even our proposed hybrid deep learning model system achieved higher accuracy and improved the UAV detection across dual vision RGB and IR. However, the varying sizes and shapes of UAVs, along with the frequency of environmental or background lighting changes, create challenges for the hybrid model. Integrating an advanced residual CNN backbone may improve detection accuracy by enhancing feature representation and facilitating more stable gradient propagation during training. The identity shortcut allows gradient magnitudes to remain stable during in-depth feature learning. Dense/UNet++ nested architectures consist of densely connected encoder–decoder bridges that utilize skip connections and retain RGB and IR features for various UAVs and shapes (Du et al., 2025). While the dual-encoder UNet protects low-level features to slow down the training of early layers, denoising and freezing later blocks adapt to changes in illumination (Wang et al., 2025a). In adverse daytime weather, the channel- and spatial-gating methods within the residual groups activate feature information and maintain performance under low-contrast IR conditions (Li and Huang, 2025). Large-kernel residual CNNs, which employ depthwise or dilated kernels with residual scaling, are beneficial for small UAV detection in cluttered RGB and IR videos. By enlarging the receptive field, these networks preserve long-range spatial context and enhance feature representation, which improves detection performance in complex environments (Liu et al., 2024a).

Swin transformers, through the use of adaptive attention transformers, prevent the fading of small target RGB and IR features at high pyramid levels (Choi et al., 2024). Moreover, the residual pre-norm layers with hierarchical windows preserve training stability and can maintain gradient health using CNN–Swin encoders for the integration of background feature information (Ding et al., 2025). Transformer-based image restoration for the task of UAV detection and tracking might be possible using frequency detection or histogram-guided attention. This method can enhance the visibility of small UAVs in both RGB and IR videos, even when the UAV is stationary, by using residual structures and pre-norms to stabilize deeper sequences of images (Gao et al., 2024; Tan et al., 2024; Wen et al., 2025a). The phase-invariant and detail-connection modules are embedded in the residual module with the structure of scaling and controlled skip-paths to keep low texture IR and RGB inputs to improve the resolution for better detection and tracking (Dan et al., 2024; Meng et al., 2025; Huang et al., 2024). Finally, vision transformers that incorporate dense residual topologies, large-kernel residuals, pre-norm transformers, and hierarchical attention with deep supervision may effectively capture cross-condition appearance shift features from both RGB and IR. It can mitigate modality imbalance in dual-vision UAV training and facilitate more stable gradient propagation during model optimization, improving robustness under challenging environmental conditions (Wen et al., 2025b; Wan et al., 2024).

6.2.2 Environmental and Background Effects in Vision Systems

The OF-GPRN system model achieved high accuracy in detecting UAVs by effectively fusing RGB and IR images and subtracting the background, even in complex environments with simi-

lar foreground elements. However, further exploration of several factors is necessary to enhance its performance. For multimodel feature integration and reliable detection, it is important to adjust modality contributions using a coarse-to-fine fusion strategy for UAV detection (Zhao et al., 2024b). Moreover, dynamic fusion approaches can adjust the RGB and IR feature fusion modalities, which might enhance the OF-GPRN's robustness during illumination variations (Zhang et al., 2025). We can enhance UAV motion estimation using more advanced optical flow methods, which can also improve UAV trajectory prediction (Cabrera et al., 2024). OF-GPRN's system model is faster with GPUs but computationally costly with CPU machines. NeuFlow has the potential to lower computational costs and enhance the efficiency of a CPU-based real-time tracking system model (Zhang et al., 2024b). Graph-based attention methods may improve spatial feature extraction and model comprehensibility. Applied graph attention networks can enhance the OF-GPRN system model for small UAV detection in cluttered environments (Zhang et al., 2025).

Another important aspect is the class imbalance in the UAV dataset, as UAVs appear sparsely in aerial imagery. Prioritizing complex images in the dataset and enhancing their representation is another crucial aspect in which robustness in complex scenarios should be a priority for high accuracy (Shi et al., 2025). Moreover, optimizing the computational cost of the OF-GPRN system model without compromising the UAV detection accuracy might be possible with pruning, compression, or multi-feature fusion networks for on-edge devices (Li et al., 2025a; Chen et al., 2024). To improve temporal consistency and operational efficiency, end-to-end multi-task learning can integrate tracking, classification, and detection in a single model (Zhang et al., 2025). The issues of reducing misalignment and improving multimodal integration in RGB and IR cross-modal alignment can be addressed using offset-guided semantic alignment with attention-based fusion (Xu et al., 2024; Qiu et al., 2025; Zhao et al., 2025). Similarly, motion-aware detection and trajectory prediction of temporal feature enhancement in UAVs can be studied, for instance, via multi-scale residual pyramids (Jiang et al., 2024; Van Lier et al., 2025). The OF-GPRN system model faces challenges due to diverse environmental conditions, occlusions, and sensor noise. It is important to explore domain adaptations and adversarial training strategies to maintain performance and enhance the model's robustness (Zhao et al., 2024b; Li et al., 2025a; Jiang et al., 2024). Moreover, integrating either optical flow-based collision detection techniques or synthetic data augmentation methods could enhance performance in multi-UAV tracking and collision prediction, overcome the data scarcity, and improve generalization. The OF-GPRN system model needs to be more attentive, visualized, interpreted, and explained for mission-critical UAV operations (Zhang et al., 2025).

6.2.3 Transmission Latency Optimization

Although our proposed A2C model for GCN assistance is capable and effectively optimizes latency while maintaining detection accuracy, it currently focuses on optimizing latency between a single UAV and the edge server. To extend the proposed model with multiple UAVs, collaborations can bring larger area coverage and can jointly schedule transmission latency optimization

and ROI selections and minimize the redundant data transfer and sharing of onboard computation. Recently researchers used a multi-agent DRL model for MEC data offloading with multi-UAV assistance, which coordinates task partitioning and trajectory planning under uncertain conditions (Wang et al., 2025b; Li et al., 2025b; Zhang et al., 2024a). Moreover, in multi-UAV beamforming and ISAC systems, the optimization of integrated sensing and communication using coordinated trajectory and beamforming design has demonstrated significant performance improvements (Wu et al., 2024b; Liu et al., 2024b; Deng et al., 2025). In the meantime, distributed algorithms and trajectory scheduling have effectively resolved multi-UAV collaborative trajectory planning in dynamic situations (Cai et al., 2024; Li et al., 2024a; Sun and Li, 2025). Adding coordination features into the CMDP model, allowing UAVs to negotiate ROI coverage and allocate detection responsibilities, would greatly enhance the importance of our research in complex aviation situations.

Another promising research novelty would be to improve the GCN-assisted A2C model computational and robustness limitations. Lightweight UAV hardware needs efficient and pruned architectures that might be possible using advanced GCN models and more lightweight mobile-based deep learning models (Lin et al., 2026; Liu et al., 2025). It is possible to significantly reduce the memory overhead and onboard computations. Moreover, GCN-assisted A2C can be extended to cloud offloading with jointly optimized computation offloading to nearby edge servers using multi-UAV MEC collaboration, which might improve the ROI cropping decisions (Chang et al., 2024).

All these approaches are important and time-based and necessary for real-world UAV onboard monitoring of no-fly zones due to the more robust and increased number of UAVs in the commercial market. Together, these improvements will not only improve the latency optimization for large-area monitoring but also bring reliability, deployment readiness, and scalability to the proposed model.

References

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. doi: 10.1109/TPAMI.2012.120.

Cemal Aker and Sinan Kalkan. Using deep networks for drone detection. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2017. doi: 10.1109/AVSS.2017.8078539.

Rubina Akter, Mohtasin Golam, Van-Sang Doan, Jae-Min Lee, and Dong-Seong Kim. Iomt-net: Blockchain integrated unauthorized uav localization using lightweight convolution neural network for internet of military things. *IEEE Internet of Things Journal*, pages 1–1, 2022. doi: 10.1109/JIOT.2022.3176310.

Sara Al-Emadi and Felwa Al-Senaid. Drone detection approach based on radio-frequency using convolutional neural network. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pages 29–34, 2020. doi: 10.1109/ICIoT48696.2020.9089489.

Sara Al-Emadi, Abdulla Al-Ali, Amr Mohammad, and Abdulaziz Al-Ali. Audio based drone detection and identification using deep learning. In *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pages 459–464, 2019. doi: 10.1109/IWCMC.2019.8766732.

Mohammad F. Al-Sa'd, Abdulla Al-Ali, Amr Mohamed, Tamer Khattab, and Aiman Erbad. Rf-based drone detection and identification using deep learning approaches: An initiative towards a large open source drone database. *Future Generation Computer Systems*, 100:86–97, 2019. ISSN 0167-739X. doi: https://doi.org/10.1016/j.future.2019.05.007. URL https://www.sciencedirect.com/science/article/pii/S0167739X18330760.

Ahmad Alhilal, Tristan Braud, Bo Han, and Pan Hui. Nebula: Reliable low-latency video transmission for mobile cloud gaming. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3407–3417, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512276. URL https://doi.org/10.1145/3485447.3512276.

Mhd Saria Allahham, Tamer Khattab, and Amr Mohamed. Deep learning for rf-based drone detection and identification: A multi-channel 1-d convolutional neural networks approach. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pages 112–117, 2020. doi: 10.1109/ICIoT48696.2020.9089657.

Anas Alsoliman, Giulio Rigoni, Marco Levorato, Cristina Pinotti, Nils Ole Tippenhauer, and Mauro Conti. Cots drone detection using video streaming characteristics. In *Proceedings of*

the 22nd International Conference on Distributed Computing and Networking, ICDCN '21, page 166–175, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389334. doi: 10.1145/3427796.3428480. URL https://doi.org/10.1145/3427796.3428480.

Anas Alsoliman, Giulio Rigoni, Davide Callegaro, Marco Levorato, Cristina M. Pinotti, and Mauro Conti. Intrusion detection framework for invasive fpv drones using video streaming characteristics. *ACM Trans. Cyber-Phys. Syst.*, 7(2), apr 2023. ISSN 2378-962X. doi: 10.1145/3579999. URL https://doi.org/10.1145/3579999.

Muhammad Waseem Ashraf, Waqas Sultani, and Mubarak Shah. Dogfight: Detecting drones from drones videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7063–7072, 2021. doi: 10.1109/CVPR46437.2021.00699.

Jeonghyun Baek, Sungjun Hong, Jisu Kim, and Euntai Kim. Efficient pedestrian detection at nighttime using a thermal camera. *Sensors*, 17(8), 2017. ISSN 1424-8220. doi: 10.3390/s17081850. URL https://www.mdpi.com/1424-8220/17/8/1850.

Sanjoy Basak, Sreeraj Rajendran, Sofie Pollin, and Bart Scheers. Drone classification from rf fingerprints using deep residual nets. In *2021 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pages 548–555, 2021. doi: 10.1109/COMSNETS51098.2021.9352891.

Sanjoy Basak, Sreeraj Rajendran, Sofie Pollin, and Bart Scheers. Combined rf-based drone detection and classification. *IEEE Transactions on Cognitive Communications and Networking*, 8 (1):111–120, 2022. doi: 10.1109/TCCN.2021.3099114.

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: speeded up robust features. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part I*, ECCV'06, page 404–417, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540338322. doi: 10.1007/11744023_32. URL https://doi.org/10.1007/11744023_32.

Dinesh Kumar Behera and Arockia Bazil Raj. Drone detection and classification using deep learning. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1012–1016, 2020. doi: 10.1109/ICICCS48265.2020.9121150.

Andrew Brock, Theodore Lim, J. M. Ritchie, and Nick Weston. Freezeout: Accelerate training by progressively freezing layers, 2017. URL https://arxiv.org/abs/1706.04983.

Liyana Adilla binti Burhanuddin, Xiaonan Liu, Yansha Deng, Ursula Challita, and András Zahemszky. Qoe optimization for live video streaming in uav-to-uav communications via deep reinforcement learning. *IEEE Transactions on Vehicular Technology*, 71(5):5358–5370, 2022. doi: 10.1109/TVT.2022.3152146.

Juan José Cabrera, Arturo Gil, Luis Payá, Antonio Santo, Oscar Reinoso, and David Rodríguez. Detection of uavs on a collision course using optical flow. In *Proceedings of the 2024 4th International Conference on Robotics and Control Engineering*, RobCE '24, page 138–144, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400716782. doi: 10.1145/3674746.3674795. URL https://doi.org/10.1145/3674746.3674795.

Qing Cai, Zheng Tang, and Chuan Liu. Joint trajectory, transmission time and power optimization for multi-uav data collecting system. *Heliyon*, 10(5):e26627, 2024. ISSN 2405-8440. doi: https://doi.org/10.1016/j.heliyon.2024.e26627. URL https://www.sciencedirect.com/science/article/pii/S2405844024026586.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58451-1. doi: 10.1007/978-3-030-58452-8_13. URL https://doi.org/10.1007/978-3-030-58452-8_13.

Adrian Carrio, Sai Vemprala, Andres Ripoll, Srikanth Saripalli, and Pascual Campoy. Drone detection using depth maps. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1034–1037, 2018. doi: 10.1109/IROS.2018.8593405.

Shu-Wei Chang, Jian-Jhih Kuo, Mong-Jen Kao, Bo-Zhong Chen, and Qian-Jing Wang. Near-optimal uav deployment for delay-bounded data collection in iot networks. In *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications*, pages 111–120, 2024. doi: 10.1109/INFOCOM52122.2024.10621135.

Bingwen Chen, Wenwei Wang, and Qianqing Qin. Robust multi-stage approach for the detection of moving target from infrared imagery. *Optical Engineering*, 51(6):067006, 2012. doi: 10.1117/1.OE.51.6.067006. URL https://doi.org/10.1117/1.OE.51.6.067006.

Yunlei Chen, Ziyan Liu, Lihui Zhang, Yingyu Wu, Qian Zhang, and Xuhui Zheng. A lightweight multi-feature fusion network for unmanned aerial vehicle infrared ray image object detection. *The Egyptian Journal of Remote Sensing and Space Sciences*, 27(2):268–276, 2024. ISSN 1110-9823. doi: https://doi.org/10.1016/j.ejrs.2024.03.001. URL https://www.sciencedirect.com/science/article/pii/S1110982324000218.

Byunggil Choi and Daegun Oh. Classification of drone type using deep convolutional neural networks based on micro- doppler simulation. In *2018 International Symposium on Antennas and Propagation (ISAP)*, pages 1–2, 2018a.

Byunggil Choi and Daegun Oh. Classification of drone type using deep convolutional neural networks based on micro-doppler simulation. In *2018 International Symposium on Antennas and Propagation (ISAP)*, pages 1–2. IEEE, 2018b.

Lammi Choi, {Won Young} Chung, and {Chan Gook} Park. Csi-net: Cnn swin transformer integrated network for infrared small target detection. *International Journal of Control, Automation and Systems*, 22(9):2899–2908, September 2024. ISSN 1598-6446. doi: 10.1007/s12555-024-0089-8. Publisher Copyright: © ICROS, KIEE and Springer 2024.

Angelo Coluccia, Gianluca Parisi, and Alessio Fascista. Detection and classification of multirotor drones in radar sensor networks: A review. *Sensors*, 20(15), 2020. ISSN 1424-8220. doi: 10.3390/s20154172. URL https://www.mdpi.com/1424-8220/20/15/4172.

Celine Craye and Salem Ardjoune. Spatio-temporal semantic segmentation for drone detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–5, 2019. doi: 10.1109/AVSS.2019.8909854.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005. doi: 10.1109/CVPR.2005.177.

Jun Dan, Tao Jin, Hao Chi, Mushui Liu, Jiawang Yu, Keying Cao, Xinjing Yang, Luo Zhao, and Haoran Xie. Pirn: Phase invariant reconstruction network for infrared image super-resolution. *Neurocomput.*, 599(C), September 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2024.128221. URL https://doi.org/10.1016/j.neucom.2024.128221.

James W. Davis and Vinay Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(2):162–182, 2007. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2006.06.010. URL https://www.sciencedirect.com/science/article/pii/S1077314206001834. Special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum.

Maarten Uijt de Haag, Chris G. Bartone, and Michael S. Braasch. Flight-test evaluation of small form-factor lidar and radar sensors for suas detect-and-avoid applications. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pages 1–11, 2016. doi: 10.1109/DASC.2016.7778108.

Chunfang Deng, Mengmeng Wang, Liang Liu, Yong Liu, and Yunliang Jiang. Extended feature pyramid network for small object detection. *IEEE Transactions on Multimedia*, 24:1968–1979, 2022. doi: 10.1109/TMM.2021.3074273.

Dan Deng, Wen Zhou, Xingwang Li, Daniel Benevides da Costa, Derrick Wing Kwan Ng, and Arumugam Nallanathan. Joint beamforming and uav trajectory optimization for covert communications in isac networks. *IEEE Transactions on Wireless Communications*, 24(2):1016–1030, 2025. doi: 10.1109/TWC.2024.3503726.

Marwan Dhuheir, Aiman Erbad, and Sinan Sabeeh. Llhr: Low latency and high reliability cnn distributed inference for resource-constrained uav swarms. In *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2023. doi: 10.1109/WCNC55385.2023.10118908.

Laurens Diels, Michiel Vlaminck, Bart De Wit, Wilfried Philips, and Hiep Luong. On the optimal mounting angle for a spinning lidar on a uav. *IEEE Sensors Journal*, 22(21):21240–21247, 2022. doi: 10.1109/JSEN.2022.3208434.

Leqi Ding, Yilin Xia, Biyun Wang, and Yunze Cai. Infrared dim small target detection algorithm based on adaptive attention transformer network. *Journal of Shanghai Jiaotong University (Science)*, 2025. ISSN 1995-8188. doi: 10.1007/s12204-025-2823-7. URL https://doi.org/10.1007/s12204-025-2823-7.

Q. Dong and Q. Zou. Visual uav detection method with online feature classification. In *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pages 429–432, 2017.

Yu Dong, Li Song, Rong Xie, and Wenjun Zhang. Ultra-low latency, stable, and scalable video transmission for free-viewpoint video services. *IEEE Transactions on Broadcasting*, 68(3):636–650, 2022. doi: 10.1109/TBC.2022.3176193.

Jedrzej Drozdowicz, Maciej Wielgo, Piotr Samczynski, Krzysztof Kulpa, Jaroslaw Krzonkalla, Maj Mordzonek, Marcin Bryl, and Zbigniew Jakielaszek. 35 ghz fmcw drone detection system. In *2016 17th International Radar Symposium (IRS)*, pages 1–4, 2016. doi: 10.1109/IRS.2016.7497351.

Xinyue Du, Ke Cheng, Jin Zhang, Yuanyu Wang, Fan Yang, Wei Zhou, and Yu Lin. Infrared small target detection algorithm based on improved dense nested u-net network. *Sensors*, 25(3), 2025. ISSN 1424-8220. doi: 10.3390/s25030814. URL https://www.mdpi.com/1424-8220/25/3/814.

Qingsong Fan, Yiting Li, Muhammet Deveci, Kaiyang Zhong, and Seifedine Kadry. Ludyolo: A novel lightweight object detection network for unmanned aerial vehicle. *Information Sciences*, 686:121366, 2025. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2024.121366. URL https://www.sciencedirect.com/science/article/pii/S0020025524012805.

Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, Sep 2004. ISSN 1573-1405. doi: 10.1023/B:VISI.0000022288.19776.77. URL https://doi.org/10.1023/B:VISI.0000022288.19776.77.

F. Fioranelli, M. Ritchie, H. Griffiths, and H. Borrion. Classification of loaded/unloaded micro-drones using multistatic radar. *Electronics Letters*, 51(22):1813–1815, 2015. doi: https://doi.org/10.1049/el.2015.3038. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/el.2015.3038.

Apostolos Galanopoulos, Jose A. Ayala-Romero, Douglas J. Leith, and George Iosifidis. Automl for video analytics with edge computing. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pages 1–10, 2021. doi: 10.1109/INFOCOM42981.2021.9488704.

Tao Gao, Yuanbo Wen, Kaihao Zhang, Jing Zhang, Ting Chen, Lidong Liu, and Wenhan Luo. Frequency-oriented efficient transformer for all-in-one weather-degraded image restoration. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3):1886–1899, 2024. doi: 10.1109/TCSVT.2023.3299324.

Anurag Ghosh, Srinivasan Iyengar, Stephen Lee, Anuj Rathore, and Venkata N Padmanabhan. React: Streaming video analytics on the edge with asynchronous cloud support. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, IoTDI '23, page 222–235, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700378. doi: 10.1145/3576842.3582385. URL https://doi.org/10.1145/3576842.3582385.

Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2018.11.017. URL https://www.sciencedirect.com/science/article/pii/S1566253517308138.

Lav Gupta, Raj Jain, and Gabor Vaszkun. Survey of important issues in uav communication networks. *IEEE Communications Surveys & Tutorials*, 18(2):1123–1152, 2016. doi: 10.1109/COMST.2015.2495297.

Jin Han, Yun-Feng Ren, Alessandro Brighente, and Mauro Conti. Rango: A novel deep learning approach to detect drones disguising from video surveillance systems. *ACM Trans. Intell. Syst. Technol.*, 15(2), feb 2024. ISSN 2157-6904. doi: 10.1145/3641282. URL https://doi.org/10.1145/3641282.

Kai Han, Yunhe Wang, Jianyuan Guo, Yehui Tang, and Enhua Wu. Vision gnn: an image is worth graph of nodes. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015. doi: 10.1109/TPAMI.2015.2389824.

Kun He, Yan Lu, and Stan Sclaroff. Local descriptors optimized for average precision. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 596–605, 2018. doi: 10.1109/CVPR.2018.00069.

Qingdong He, Zhengning Wang, Hao Zeng, Yi Zeng, and Yijun Liu. Svga-net: Sparse voxel-graph attention network for 3d object detection from point clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 870–878, 2022.

Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2883–2892, 2019. doi: 10.1109/CVPR.2019.00300.

Aaron Hitchcock and Kelvin Sung. Multi-view augmented reality with a drone. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, VRST '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360869. doi: 10.1145/3281505.3283397. URL https://doi.org/10.1145/3281505.3283397.

Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. Deeply supervised salient object detection with short connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):815–828, 2019. doi: 10.1109/TPAMI.2018.2815688.

Simon Hradecky. Lam b737 at tete on jan 5th 2017, radome structural failure, 2017. URL https://www.avherald.com/h?article=4a319157. Initially reported as a suspected drone strike; subsequent regulator statements indicated likely structural failure of the radome.

Y. Hu, X. Wu, G. Zheng, and X. Liu. Object detection of uav for anti-uav based on improved yolo v3. In *2019 Chinese Control Conference (CCC)*, pages 8386–8390, 2019.

Feng Huang, Yunxiang Li, Xiaojing Ye, and Jing Wu. Infrared image super-resolution network utilizing the enhanced transformer and u-net. *Sensors*, 24(14), 2024. ISSN 1424-8220. doi: 10.3390/s24144686. URL https://www.mdpi.com/1424-8220/24/14/4686.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 646–661, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

Fortune Business Insights. Unmanned aerial vehicle (uav) market size, share & covid-19 impact analysis, by type (fixed-wing, rotary-wing, hybrid), by system (airframe, payloads, avionics, propulsion), by technology (remotely operated, semi-autonomous, fully autonomous), by application (military, commercial, government & law enforcement, consumer), and regional forecast, 2023-2030, 2024. URL https://www.fortunebusinessinsights.com/industry-reports/unmanned-aerial-vehicle-uav-market-101603. Accessed: November 28, 2024.

Brian K. S. Isaac-Medina, Matt Poyser, Daniel Organisciak, Chris G. Willcocks, Toby P. Breckon, and Hubert P. H. Shum. Unmanned aerial vehicle visual detection and tracking using deep neural networks: A performance benchmark. In *2021 IEEE/CVF International Conference on*

Computer Vision Workshops (ICCVW), pages 1223–1232, 2021a. doi: 10.1109/ICCVW54120.2021.00142.

Brian K. S. Isaac-Medina, Matt Poyser, Daniel Organisciak, Chris G. Willcocks, Toby P. Breckon, and Hubert P. H. Shum. Unmanned aerial vehicle visual detection and tracking using deep neural networks: A performance benchmark. *arXiv*, 2021b.

Mohammed Jahangir and Chris Baker. Robust detection of micro-uas drones with l-band 3-d holographic radar. In *2016 Sensor Signal Processing for Defence (SSPD)*, pages 1–5, 2016. doi: 10.1109/SSPD.2016.7590610.

Mohammed Jahangir and Chris J Baker. Extended dwell doppler characteristics of birds and micro-uas at l-band. In *2017 18th International Radar Symposium (IRS)*, pages 1–10, 2017. doi: 10.23919/IRS.2017.8008144.

Mohammed Jahangir, Chris J Baker, and Gordon A Oswald. Doppler characteristics of micro-drones with l-band multibeam staring radar. In *2017 IEEE Radar Conference (RadarConf)*, pages 1052–1057, 2017. doi: 10.1109/RADAR.2017.7944360.

Di Jia, Kai Wang, ShunLi Luo, TianYu Liu, and Ying Liu. Braft: Recurrent all-pairs field transforms for optical flow based on correlation blocks. *IEEE Signal Processing Letters*, 28:1575–1579, 2021. doi: 10.1109/LSP.2021.3099076.

Changjun Jiang, Yu Fang, Peihai Zhao, and John Panneerselvam. Intelligent uav identity authentication and safety supervision based on behavior modeling and prediction. *IEEE Transactions on Industrial Informatics*, 16(10):6652–6662, 2020. doi: 10.1109/TII.2020.2966758.

Lingjie Jiang, Baoxi Yuan, Jiawei Du, Boyu Chen, Hanfei Xie, Juan Tian, and Ziqi Yuan. Mff-sodnet: Multiscale feature fusion small object detection network for uav aerial images. *IEEE Transactions on Instrumentation and Measurement*, 73:1–14, 2024. doi: 10.1109/TIM.2024.3381272.

Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Qixiang Ye, Jianbin Jiao, Zhenjun Han, et al. Anti-uav: a large-scale benchmark for vision-based uav tracking. *IEEE Transactions on Multimedia*, 2021.

Ren Jin, Jiaqi Jiang, Yuhua Qi, Defu Lin, and Tao Song. Drone detection and pose estimation using relational graph networks. *Sensors*, 19(6), 2019. ISSN 1424-8220. doi: 10.3390/s19061479. URL https://www.mdpi.com/1424-8220/19/6/1479.

Honggu Kang, Jingon Joung, Jinyoung Kim, Joonhyuk Kang, and Yong Soo Cho. Protect your sky: A survey of counter unmanned aerial vehicle systems. *IEEE Access*, 8:168671–168710, 2020. doi: 10.1109/ACCESS.2020.3023473.

Muhammad Asif Khan, Emna Baccour, Zina Chkirbene, Aiman Erbad, Ridha Hamila, Mounir Hamdi, and Moncef Gabbouj. A survey on mobile edge computing for video streaming: Opportunities and challenges. *IEEE Access*, 10:120514–120550, 2022. doi: 10.1109/ACCESS.2022.3220694.

Naveed Khan, Ayaz Ahmad, Abdul Wakeel, Zeeshan Kaleem, Bushra Rashid, and Waqas Khalid. Efficient uavs deployment and resource allocation in uav-relay assisted public safety networks for video transmission. *IEEE Access*, 12:4561–4574, 2024. doi: 10.1109/ACCESS.2024.3350138.

Byung Kwan Kim, Hyun-Seong Kang, and Seong-Ook Park. Drone classification using convolutional neural networks with merged doppler images. *IEEE Geoscience and Remote Sensing Letters*, 14(1):38–42, 2017a. doi: 10.1109/LGRS.2016.2624820.

Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee. Residual lstm: Design of a deep recurrent architecture for distant speech recognition. In Francisco Lacerda, editor, *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 1591–1595. ISCA, 2017b. URL http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0477.html.

Jongha Kim, Jinheon Baek, and Sung Ju Hwang. Object detection in aerial images with uncertainty-aware graph network. In *Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, page 521–536, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-25068-2. doi: 10.1007/978-3-031-25069-9_34. URL https://doi.org/10.1007/978-3-031-25069-9_34.

Aybora Koksal, Kutalmis Gokalp Ince, and A. Aydin Alatan. Effect of annotation errors on drone detection with yolov3. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4439–4447, 2020. doi: 10.1109/CVPRW50498.2020.00523.

Harrison Kurunathan, Hailong Huang, Kai Li, Wei Ni, and Ekram Hossain. Machine learning-aided operations and communications of unmanned aerial vehicles: A contemporary survey. *IEEE Communications Surveys & Tutorials*, 26(1):496–533, 2024. doi: 10.1109/COMST.2023.3312221.

Asif Ali Laghari, Awais Khan Jumani, Rashid Ali Laghari, Hang Li, Shahid Karim, and Abudllah Ayub Khan. Unmanned aerial vehicles advances in object detection and communication security review. *Cognitive Robotics*, 4:128–141, 2024. ISSN 2667-2413. doi: https://doi.org/10.1016/j.cogr.2024.07.002. URL https://www.sciencedirect.com/science/article/pii/S2667241324000090.

Ying-Chih Lai and Zong-Ying Huang. Detection of a moving uav based on deep learning-based distance estimation. *Remote Sensing*, 12(18), 2020. ISSN 2072-4292. doi: 10.3390/rs12183035. URL https://www.mdpi.com/2072-4292/12/18/3035.

Latvian Public Broadcasting LSM. Latvia's runaway drone was flying unlicensed, 2020. URL https://eng.lsm.lv/article/economy/transport/latvias-runaway-drone-was-flying-unlicensed.a361222/. Airspace around Rīga affected after a rogue drone went out of control on May 2, 2020.

Latvian Public Broadcasting LSM. Planes redirected from rīga due to drone presence, 2025. URL https://eng.lsm.lv/article/economy/transport/14.01.2025-planes-redirected-from-riga-monday-due-to-drone-presence.a583539/. Three flights diverted on Jan 13, 2025 after a drone was detected near the landing trajectory.

Dongkyu Lee, Woong Gyu La, and Hwangnam Kim. Drone detection and identification system using artificial intelligence. In *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1131–1133, 2018. doi: 10.1109/ICTC.2018.8539442.

Alex Leykin, Yang Ran, and Riad Hammoud. Thermal-visible video fusion for moving target tracking and pedestrian classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. doi: 10.1109/CVPR.2007.383444.

Bin Li, Shi Qiu, Wei Jiang, Wei Zhang, Mingnan Le, and Kun Wei. A uav detection and tracking algorithm based on image feature super-resolution. *Wirel. Commun. Mob. Comput.*, 2022, January 2022a. ISSN 1530-8669. doi: 10.1155/2022/6526684. URL https://doi.org/10.1155/2022/6526684.

Bin Li, Rongrong Yang, Lei Liu, Junyi Wang, Ning Zhang, and Mianxiong Dong. Robust computation offloading and trajectory optimization for multi-uav-assisted mec: A multiagent drl approach. *IEEE Internet of Things Journal*, 11(3):4775–4786, 2024a. doi: 10.1109/JIOT.2023.3300718.

Chen Li, Rui Zhao, Zeyu Wang, Huiying Xu, and Xinzhong Zhu. Remdet: rethinking efficient model design for uav object detection. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025a. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i5.32490. URL https://doi.org/10.1609/aaai.v39i5.32490.

Guohao Li, Matthias Mueller, Guocheng Qian, Itzel Carolina Delgadillo Perez, Abdulellah Abualshour, Ali Kassem Thabet, and Bernard Ghanem. Deepgcns: Making gcns go as deep as cnns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021a. doi: 10.1109/TPAMI.2021.3074057.

Hangyu Li, Fahui Miao, and Xiaojun Mei. Facilitating multi-uavs application for rescue in complex 3d sea wind offshore environment: A scalable multi-uavs collaborative path planning method based on improved coatis optimization algorithm. *Ocean Engineering*, 324:120701, 2025b. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2025.120701. URL https://www.sciencedirect.com/science/article/pii/S0029801825004160.

Jing Li, Dong Hye Ye, Mathias Kolsch, Juan P. Wachs, and Charles A Bouman. Fast and robust uav to uav detection and tracking from video. *IEEE Transactions on Emerging Topics in Computing*, pages 1–1, 2021b. doi: 10.1109/TETC.2021.3104555.

Kai Li, Razvan Christian Voicu, Salil S. Kanhere, Wei Ni, and Eduardo Tovar. Energy efficient legitimate wireless surveillance of uav communications. *IEEE Transactions on Vehicular Technology*, 68(3):2283–2293, 2019. doi: 10.1109/TVT.2019.2890999.

Kai Li, Wei Ni, Alam Noor, and Mohsen Guizani. Employing intelligent aerial data aggregators for the internet of things: Challenges and solutions. *IEEE Internet of Things Magazine*, 5(1):136–141, 2022b. doi: 10.1109/IOTM.001.2100161.

Kai Li, Wei Ni, Xin Yuan, Alam Noor, and Abbas Jamalipour. Deep graph-based reinforcement learning for joint cruise control and task offloading for aerial edge internet-of-things (edgeiot). *IEEE Internet of Things Journal*, pages 1–1, 2022c. doi: 10.1109/JIOT.2022.3182119.

Kai Li, Wei Ni, Xin Yuan, Alam Noor, and Abbas Jamalipour. Exploring graph neural networks for joint cruise control and task offloading in uav-enabled mobile edge computing. In *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, pages 1–6, 2023. doi: 10.1109/VTC2023-Spring57618.2023.10200966.

Long Li, Gaochao Xu, Ziqi Liu, Xu Xu, Xianqiu Meng, and Xiangyu Meng. Multiobjective optimization of energy efficiency and fairness in aav-assisted wireless powered mec systems: A drl-based approach. *IEEE Internet of Things Journal*, 12(14):28758–28775, 2025c. doi: 10.1109/JIOT.2025.3566865.

Sihan Li and Juhua Huang. Resgdanet: An efficient residual group attention neural network for medical image classification. *Applied Sciences*, 15(5), 2025. ISSN 2076-3417. doi: 10.3390/app15052693. URL https://www.mdpi.com/2076-3417/15/5/2693.

Wei Li, Dequan Zheng, Tiejun Zhao, and Mengda Yang. An effective approach to pedestrian detection in thermal imagery. In *2012 8th International Conference on Natural Computation*, pages 325–329, 2012. doi: 10.1109/ICNC.2012.6234621.

Yangang Li, Qi Li, Jie Pan, Ying Zhou, Hongliang Zhu, Hongwei Wei, and Chong Liu. Sod-yolo: Small-object-detection algorithm based on improved yolov8 for uav images. *Remote Sensing*, 16(16), 2024b. ISSN 2072-4292. doi: 10.3390/rs16163057. URL https://www.mdpi.com/2072-4292/16/16/3057.

Zheng Li, Xiaocong Du, and Yu Cao. Gar: Graph assisted reasoning for object detection. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1284–1293, 2020. doi: 10.1109/WACV45572.2020.9093559.

Conggong Lin, Yushi Zhang, and Guodong Chen. Lightweight underwater debris detection model based on improved rt-detr. *Marine Pollution Bulletin*, 222:118537, 2026. ISSN 0025-326X. doi: https://doi.org/10.1016/j.marpolbul.2025.118537. URL https://www.sciencedirect.com/science/article/pii/S0025326X25010124.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. doi: 10.1109/CVPR.2017.106.

Huiying Liu, Jiarong Wang, Weijun Zhong, Haitao Nie, Xiaotong Deng, Jiaqi Sun, Ming Zhu, and Ming Wei. Res-lk-slr: A residual network based on large kernels and shapelet-level representations for spatial infrared spot target discrimination. *Remote Sensing*, 16(24), 2024a. ISSN 2072-4292. doi: 10.3390/rs16244624. URL https://www.mdpi.com/2072-4292/16/24/4624.

Saichao Liu, Geng Sun, Jiahui Li, Shuang Liang, Qingqing Wu, Pengfei Wang, and Dusit Niyato. Uav-enabled collaborative beamforming via multi-agent deep reinforcement learning. *IEEE Transactions on Mobile Computing*, 23(12):13015–13032, 2024b. doi: 10.1109/TMC.2024.3419915.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018. doi: 10.1109/CVPR.2018.00913.

Yicheng Liu, Luchuan Liao, Hao Wu, Jing Qin, Ling He, Gang Yang, Han Zhang, and Jing Zhang. Trajectory and image-based detection and identification of uav. *Vis. Comput.*, 37(7):1769–1780, July 2021. ISSN 0178-2789. doi: 10.1007/s00371-020-01937-y. URL https://doi.org/10.1007/s00371-020-01937-y.

Zhichao Liu and Yi Jiang. Design and implementation for a uav-based streaming media system. *Ad Hoc Networks*, 156:103443, 2024. ISSN 1570-8705. doi: https://doi.org/10.1016/

j.adhoc.2024.103443. URL https://www.sciencedirect.com/science/article/pii/S1570870524000544.

Zhilong Liu, Jiayi Zhang, Yong Zeng, and Bo Ai. Energy-efficient multi-agent reinforcement learning for uav trajectory optimization in cell-free massive mimo networks. *IEEE Transactions on Wireless Communications*, 24(7):5917–5930, 2025. doi: 10.1109/TWC.2025.3550266.

Zihao Liu, Haiqin Xu, Yihong Zhang, Zhouyi Xu, Sen Wu, and Di Zhu. A real-time detection drone algorithm based on instance semantic segmentation. In *Proceedings of the 3rd International Conference on Video and Image Processing*, ICVIP '19, page 36–41, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450376822. doi: 10.1145/3376067.3376098. URL https://doi.org/10.1145/3376067.3376098.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004. ISSN 0920-5691. doi: 10.1023/B:VISI.0000029664.99615.94. URL https://doi.org/10.1023/B:VISI.0000029664.99615.94.

Vasileios Magoulianitis, Dimitrios Ataloglou, Anastasios Dimou, Dimitrios Zarpalas, and Petros Daras. Does deep super-resolution enhance uav detection? In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2019. doi: 10.1109/AVSS.2019.8909865.

Antonino Masaracchia, Yijiu Li, Khoi Khac Nguyen, Cheng Yin, Saeed R. Khosravirad, Daniel Benevides Da Costa, and Trung Q. Duong. Uav-enabled ultra-reliable low-latency communications for 6g: A comprehensive survey. *IEEE Access*, 9:137338–137352, 2021. doi: 10.1109/ACCESS.2021.3117902.

Ziang Meng, Kaiqi Han, Yuchun He, Yuhan He, Xingyuan Li, and Yang Zou. Modeling detail feature connections for infrared image enhancement. *Neurocomput.*, 639(C), August 2025. ISSN 0925-2312. doi: 10.1016/j.neucom.2025.130200. URL https://doi.org/10.1016/j.neucom.2025.130200.

Jiansong Miao, Shanling Bai, Shahid Mumtaz, Qian Zhang, and Junsheng Mu. Utility-oriented optimization for video streaming in uav-aided mec network: A drl approach. *IEEE Transactions on Green Communications and Networking*, 8(2):878–889, 2024. doi: 10.1109/TGCN.2024.3352173.

P. Molchanov, K. Egiazarian, J. Astola, R. I. A. Harmanny, and J. J. M. de Wit. Classification of small uavs and birds by micro-doppler signatures. In *2013 European Radar Conference*, pages 172–175, 2013.

M. Nalamati, A. Kapoor, M. Saqib, N. Sharma, and M. Blumenstein. Drone detection in long-range surveillance videos. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2019.

Jeffrey A. Nanzer and Victor C. Chen. Microwave interferometric and doppler radar measurements of a uav. In *2017 IEEE Radar Conference (RadarConf)*, pages 1628–1633, 2017. doi: 10.1109/RADAR.2017.7944468.

T. Ning, W. Wu, and J. Zhang. Small object detection based on yolov8 in uav perspective. *Pattern Analysis and Applications*, 27:103, 2024. doi: 10.1007/s10044-024-01323-7.

Alam Noor, Kai Li, Adel Ammar, Anis Koubaa, Bilel Benjdira, and Eduardo Tovar. A hybrid deep learning model for uavs detection in day and night dual visions. In *2021 IEEE Third*

International Conference on Cognitive Machine Intelligence (CogMI), pages 221–231, 2021. doi: 10.1109/CogMI52975.2021.00037.

Jihun Park, Dae Hoe Kim, Young Sook Shin, and Sang-ho Lee. A comparison of convolutional object detectors for real-time drone tracking using a ptz camera. In *2017 17th International Conference on Control, Automation and Systems (ICCAS)*, pages 696–699, 2017. doi: 10.23919/ICCAS.2017.8204318.

Devarsh Patel, Sarthak Patel, and Megh Patel. Application of image-to-image translation in improving pedestrian detection. In Manjaree Pandit, M. K. Gaur, and Sandeep Kumar, editors, *Artificial Intelligence and Sustainable Computing*, pages 471–482, Singapore, 2023. Springer Nature Singapore. ISBN 978-981-99-1431-9.

Junkai Peng, Changwen Zheng, Pin Lv, Tianyu Cui, Ye Cheng, and Si Lingyu. Using images rendered by pbrt to train faster r-cnn for uav detection. 2018.

Jordi Pont-Tuset, Pablo Arbeláez, Jonathan T. Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):128–140, 2017. doi: 10.1109/TPAMI.2016.2537320.

UK POST. Misuse of civilian drones. Postnote 610, UK Parliamentary Office of Science and Technology, 2020. URL https://researchbriefings.files.parliament.uk/documents/POST-PN-0610/POST-PN-0610.pdf. Includes a summary of the December 2018 Gatwick disruption (1,000 flights; 36 hours; >140,000 passengers).

Shengxiang Qi, Wei Zhang, and Guojing Xu. Detecting consumer drones from static infrared images by fast-saliency and hog descriptor. In *Proceedings of the 4th International Conference on Communication and Information Processing*, ICCIP '18, page 62–66, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450365345. doi: 10.1145/3290420.3290426. URL https://doi.org/10.1145/3290420.3290426.

Yuxuan Qiu, Liyang Wang, Wei Song, Jiawei Liu, Zhiping Shi, and Na Jiang. Advancing visible-infrared person re-identification: Synergizing visual-textual reasoning and cross-modal feature alignment. *IEEE Transactions on Information Forensics and Security*, 20:2184–2196, 2025. doi: 10.1109/TIFS.2025.3539946.

Chengyi Qu, Paulo Drefahl, Wenbin Guo, and Hong Wang. Autonomous video transmission and air-to-ground coordination in uav-swarm-aided disaster response platform. In Vincent G. Duffy, editor, *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*, pages 339–355, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-61060-8.

K Venkata Satya Reddy, Venkata Ramana Molabanti, Sai Tulasi Prasad Dumpala, and Usha Rani. Nelakuditi. Yolov3 based real time drone detection for counter drone system. In *2023 IEEE 3rd International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET)*, pages 1–5, 2023. doi: 10.1109/TEMSMET56707.2023.10149935.

Khosro Rezaee, Mohammad R. Khosravi, and Maryam Saberi Anari. Deep-transfer-learning-based abnormal behavior recognition using internet of drones for crowded scenes. *IEEE Internet of Things Magazine*, 5(2):41–44, 2022. doi: 10.1109/IOTM.001.2100138.

Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Flying objects detection from a single moving camera. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4128–4136, 2015. doi: 10.1109/CVPR.2015.7299040.

Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Detecting flying objects using a single moving camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):879–892, 2017. doi: 10.1109/TPAMI.2016.2564408.

Luana Ruiz, Fernando Gama, and Alejandro Ribeiro. Graph neural networks: Architectures, stability, and transferability. *Proceedings of the IEEE*, 109(5):660–682, 2021. doi: 10.1109/JPROC.2021.3055400.

Stamatios Samaras, Eleni Diamantidou, Dimitrios Ataloglou, Nikos Sakellariou, Anastasios Vafeiadis, Vasilis Magoulianitis, Antonios Lalas, Anastasios Dimou, Dimitrios Zarpalas, Konstantinos Votis, Petros Daras, and Dimitrios Tzovaras. Deep learning on multi sensor data for counter uav applications—a systematic review. *Sensors*, 19(22), 2019. ISSN 1424-8220. doi: 10.3390/s19224837. URL https://www.mdpi.com/1424-8220/19/22/4837.

Muhammad Saqib, Sultan Daud Khan, Nabin Sharma, and Michael Blumenstein. A study on detecting drones using deep convolutional neural networks. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–5, 2017. doi: 10.1109/AVSS.2017.8078541.

Boban Sazdić-Jotić, Ivan Pokrajac, Jovan Bajčetić, Boban Bondžulić, and Danilo Obradović. Single and multiple drones detection and identification using rf based deep learning algorithm. *Expert Systems with Applications*, 187:115928, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.115928. URL https://www.sciencedirect.com/science/article/pii/S0957417421012823.

Arne Schumann, Lars Sommer, Johannes Klatte, Tobias Schuchert, and Jürgen Beyerer. Deep cross-domain flying object classification for robust uav detection. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2017. doi: 10.1109/AVSS.2017.8078558.

Ulzhalgas Seidaliyeva, Daryn Akhmetov, Lyazzat Ilipbayeva, and Eric T. Matson. Real-time and accurate drone detection in a video with a static background. *Sensors*, 20(14), 2020a. ISSN 1424-8220. doi: 10.3390/s20143856. URL https://www.mdpi.com/1424-8220/20/14/3856.

Ulzhalgas Seidaliyeva, Manal Alduraibi, Lyazzat Ilipbayeva, and Akhan Almagambetov. Detection of loaded and unloaded uav using deep neural network. In *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*, pages 490–494, 2020b. doi: 10.1109/IRC.2020.00093.

Ulzhalgas Seidaliyeva, Manal Alduraibi, Lyazzat Ilipbayeva, and Nurzhigit Smailov. Deep residual neural network-based classification of loaded and unloaded uav images. In *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*, pages 465–469, 2020c. doi: 10.1109/IRC.2020.00088.

Samira Shackle. The mystery of the gatwick drone, 2020. URL https://www.theguardian.com/uk-news/2020/dec/01/the-mystery-of-the-gatwick-drone. Context/features piece on the 2018 Gatwick incident.

Mohit K. Sharma, Chen-Feng Liu, Ibrahim Farhat, Nassim Sehad, Wassim Hamidouche, and Merouane Debbah. Uav immersive video streaming: A comprehensive survey, benchmarking, and open challenges, 2023. URL https://arxiv.org/abs/2311.00082.

Mohit K. Sharma, Ibrahim Farhat, Chen-Feng Liu, Nassim Sehad, Wassim Hamidouche, and Mérouane Debbah. Real-time immersive aerial video streaming: A comprehensive survey, benchmarking, and open challenges. *IEEE Open Journal of the Communications Society*, 5: 5680–5705, 2024. doi: 10.1109/OJCOMS.2024.3455763.

Lin Shi, Ishtiaq Ahmad, YuJing He, and KyungHi Chang. Hidden markov model based drone sound recognition using mfcc technique in practical noisy environments. *Journal of Communications and Networks*, 20(5):509–518, 2018. doi: 10.1109/JCN.2018.000075.

Shuohao Shi, Qiang Fang, and Xin Xu. Joint optimization loss function for tiny object detection in remote sensing images. *Remote Sensing*, 17(8), 2025. ISSN 2072-4292. doi: 10.3390/rs17081476. URL https://www.mdpi.com/2072-4292/17/8/1476.

Weijing Shi and Raj Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1708–1716, 2020. doi: 10.1109/CVPR42600.2020.00178.

Subroto Singha and Burchan Aydin. Automated drone detection using yolov4. *Drones*, 5(3), 2021. ISSN 2504-446X. doi: 10.3390/drones5030095. URL https://www.mdpi.com/2504-446X/5/3/95.

Lars Sommer, Arne Schumann, Thomas Müller, Tobias Schuchert, and Jürgen Beyerer. Flying object detection for automatic uav recognition. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2017. doi: 10.1109/AVSS.2017.8078557.

Congxi Song, Biao Han, Xiaolan Ji, Yahui Li, and Jinshu Su. Ai-driven multipath transmission: Empowering uav-based live streaming. *IEEE Network*, 38(2):202–210, 2024. doi: 10.1109/MNET.2023.3321521.

Ala Souissi, Hajer Fradi, and Panagiotis Papadakis. Leveraging event streams with deep reinforcement learning for end-to-end uav tracking, 2024. URL https://arxiv.org/abs/2410.14685.

Paul Stoica, Sanjoy Basak, Cristian Molder, and Bart Scheers. Review of counter-uav solutions based on the detection of remote control communication. In *2020 13th International Conference on Communications (COMM)*, pages 233–238, 2020. doi: 10.1109/COMM48946.2020.9142017.

Han-Sok Suh, Jian Meng, Ty Nguyen, Vijay Kumar, Yu Cao, and Jae-Sun Seo. Algorithm-hardware co-optimization for energy-efficient drone detection on resource-constrained fpga. *ACM Trans. Reconfigurable Technol. Syst.*, 16(2), may 2023. ISSN 1936-7406. doi: 10.1145/3583074. URL https://doi.org/10.1145/3583074.

Chao Sun and Zhihua Li. Multi-uav-assisted mec offloading-optimization method on deep reinforcement learning. *International Journal on Semantic Web and Information Systems*, 21 (1), 2025. ISSN 1552-6283. doi: https://doi.org/10.4018/IJSWIS.368839. URL https://www.sciencedirect.com/science/article/pii/S1552628325000201.

Fuchun Sun, Tao Kong, Wenbing Huang, Chuanqi Tan, Bin Fang, and Huaping Liu. Feature pyramid reconfiguration with consistent loss for object detection. *IEEE Transactions on Image Processing*, 28(10):5041–5051, 2019. doi: 10.1109/TIP.2019.2917781.

Han Sun, Jian Yang, Jiaquan Shen, Dong Liang, Liu Ning-Zhong, and Huiyu Zhou. Tib-net: Drone detection network with tiny iterative backbone. *IEEE Access*, 8:130697–130707, 2020. doi: 10.1109/ACCESS.2020.3009518.

Bilal Taha and Abdulhadi Shoufan. Machine learning-based drone detection and classification: State-of-the-art in research. *IEEE Access*, 7:138669–138682, 2019. doi: 10.1109/ACCESS.2019.2942944.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10778–10787, 2020. doi: 10.1109/CVPR42600.2020.01079.

Zhentao Tan, Yue Wu, Qiankun Liu, Qi Chu, Le Lu, Jieping Ye, and Nenghai Yu. Exploring the application of large-scale pre-trained models on adverse weather removal. *IEEE Transactions on Image Processing*, 33:1683–1698, 2024. doi: 10.1109/TIP.2024.3368961.

Michael Teutsch, Thomas Mueller, Marco Huber, and Juergen Beyerer. Low resolution person detection with a moving thermal infrared camera by hot spot classification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 209–216, 2014. doi: 10.1109/CVPRW.2014.40.

Ziwei Tian, Jie Huang, Yang Yang, and Weiying Nie. The drone detection based on improved yolov5. In *Proceedings of the 2022 5th Artificial Intelligence and Cloud Computing Conference*, AICCC '22, page 83–89, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398749. doi: 10.1145/3582099.3582113. URL https://doi.org/10.1145/3582099.3582113.

Mikko Uitto and Antti Heikkinen. Evaluation of live video streaming performance for low latency use cases in 5g. In *2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, pages 431–436, 2021. doi: 10.1109/EuCNC/6GSummit51104.2021.9482605.

Fath U Min Ullah, Mohammad S. Obaidat, Amin Ullah, Khan Muhammad, Mohammad Hijji, and Sung Wook Baik. A comprehensive review on vision-based violence detection in surveillance videos. *ACM Comput. Surv.*, 55(10), feb 2023. ISSN 0360-0300. doi: 10.1145/3561971. URL https://doi.org/10.1145/3561971.

Hikmat Ullah, Yaqin Zhao, Longwen Wu, Alam Noor, and Liang Zhao. Multi-modal medical image fusion technique to improve glioma classification accuracy. In *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, pages 321–325, 2021. doi: 10.1109/ICSIP52628.2021.9689018.

Oytun Ulutan, A S M Iftekhar, and B. S. Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13614–13623, 2020. doi: 10.1109/CVPR42600.2020.01363.

Eren Unlu, Emmanuel Zenou, Nicolas Riviere, and Paul-Edouard Dupouy. Deep learning-based strategies for the detection and tracking of drones using several cameras. *IPSJ Transactions on Computer Vision and Applications*, 11(1):7, 2019. ISSN 1882-6695. doi: 10.1186/s41074-019-0059-x. URL https://doi.org/10.1186/s41074-019-0059-x.

Michel Van Lier, Martin Van Leeuwen, Bastian Van Manen, Leo Kampmeijer, and Nicolas Boehrer. Evaluation of Spatio-Temporal Small Object Detection in Real-World Adverse Weather Conditions . In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 786–797, Los Alamitos, CA, USA, March 2025. IEEE Computer Society. doi: 10.1109/WACVW65960.2025.00094. URL https://doi.ieeecomputersociety.org/10.1109/WACVW65960.2025.00094.

Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, pages 705–718, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-88693-8.

Matouš Vrba, Daniel Heřt, and Martin Saska. Onboard marker-less detection and localization of non-cooperating drones for their safe interception by an autonomous aerial system. *IEEE Robotics and Automation Letters*, 4(4):3402–3409, 2019. doi: 10.1109/LRA.2019.2927130.

Amy R. Wagoner, Daniel K. Schrader, and Eric T. Matson. Survey on detection and tracking of uavs using computer vision. In *2017 First IEEE International Conference on Robotic Computing (IRC)*, pages 320–325, 2017. doi: 10.1109/IRC.2017.15.

Yecong Wan, Mingwen Shao, Yuanshuo Cheng, and Wangmeng Zuo. Image all-in-one adverse weather removal via dynamic model weights generation. *Know.-Based Syst.*, 302(C), October 2024. ISSN 0950-7051. doi: 10.1016/j.knosys.2024.112324. URL https://doi.org/10.1016/j.knosys.2024.112324.

Chengyang Wang and Caiming Zhong. Adaptive feature pyramid networks for object detection. *IEEE Access*, 9:107024–107032, 2021. doi: 10.1109/ACCESS.2021.3100369.

Jiawei Wang, Tianyu Song, Yan Zhang, Shengmin Wang, Wenhui Lin, Zhigang Wang, and Yuanyuan Qiao. Decoupled teacher for semi-supervised drone detection. In *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*, pages 1869–1873, 2022. doi: 10.1109/ICCC56324.2022.10065897.

Minghe Wang, Pan Yuan, Su Qiu, Weiqi Jin, Li Li, and Xia Wang. Dual-encoder unet-based narrowband uncooled infrared imaging denoising network. *Sensors*, 25(5), 2025a. ISSN 1424-8220. doi: 10.3390/s25051476. URL https://www.mdpi.com/1424-8220/25/5/1476.

Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11531–11539, 2020. doi: 10.1109/CVPR42600.2020.01155.

Rui Wang, Kaitao Meng, and Deshi Li. Multi-uav collaborative trajectory planning for seamless data collection and transmission. In *2025 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–7, 2025b. doi: 10.1109/WCNC61545.2025.10978431.

Weihong Wang, Jian Zhang, and Chunhua Shen. Improved human detection and classification in thermal images. In *2010 IEEE International Conference on Image Processing*, pages 2313–2316, 2010. doi: 10.1109/ICIP.2010.5649946.

Xu Wang, Zheng Yang, Jiahang Wu, Yi Zhao, and Zimu Zhou. Edgeduet: Tiling small object detection for edge assisted autonomous mobile vision. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pages 1–10, 2021a. doi: 10.1109/INFOCOM42981.2021.9488843.

Yongxin Wang, Kris Kitani, and Xinshuo Weng. Joint object detection and multi-object tracking with graph neural networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13708–13715, 2021b. doi: 10.1109/ICRA48506.2021.9561110.

Yuqin Wang, Jingce Xu, and Wen Ji. A feature-based video transmission framework for visual iot in fog computing systems. In *2019 ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS)*, pages 1–8, 2019. doi: 10.1109/ANCS.2019.8901872.

Daniel Tan Wei Xun, Yoke Lin Lim, and Sutthiphong Srigrarom. Drone detection using yolov3 with transfer learning on nvidia jetson tx2. In *2021 Second International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP)*, pages 1–6, 2021. doi: 10.1109/ICA-SYMP50206.2021.9358449.

Yang Wen, Anyu Lai, Bo Qian, Wuzhen Shi, and Wenming Cao. Multi-weather image restoration via histogram-based transformer feature enhancement. *The Visual Computer*, 2025a. ISSN 1432-2315. doi: 10.1007/s00371-025-04085-3. URL https://doi.org/10.1007/s00371-025-04085-3.

Yuanbo Wen, Tao Gao, Ziqi Li, Jing Zhang, Kaihao Zhang, and Ting Chen. All-in-one weather-degraded image restoration via adaptive degradation-aware self-prompting model. *IEEE Transactions on Multimedia*, 27:3343–3355, 2025b. doi: 10.1109/TMM.2025.3535316.

Xinshuo Weng, Yongxin Wang, Yunze Man, and Kris M. Kitani. Gnn3dmot: Graph neural network for 3d multi-object tracking with 2d-3d multi-feature learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6498–6507, 2020. doi: 10.1109/CVPR42600.2020.00653.

Dan Wu, Liming Wang, Meiyan Liang, Yunpeng Kang, Qi Jiao, Yajun Cheng, and Jian Li. Uav-assisted real-time video transmission for vehicles: A soft actor–critic drl approach. *IEEE Internet of Things Journal*, 11(8):14710–14726, 2024a. doi: 10.1109/JIOT.2023.3343590.

Fanyi Wu, Hongliang Zhang, Jianjun Wu, Zhu Han, H. Vincent Poor, and Lingyang Song. Uav-to-device underlay communications: Age of information minimization by multi-agent deep reinforcement learning. *IEEE Transactions on Communications*, 69(7):4461–4475, 2021. doi: 10.1109/TCOMM.2021.3065135.

Jiyan Wu, Bo Cheng, Ming Wang, and Junliang Chen. Energy-efficient bandwidth aggregation for delay-constrained video over heterogeneous wireless networks. *IEEE Journal on Selected Areas in Communications*, 35(1):30–49, 2017. doi: 10.1109/JSAC.2016.2632599.

Jiyan Wu, Rui Tan, and Ming Wang. Streaming high-definition real-time video to mobile devices with partially reliable transfer. *IEEE Transactions on Mobile Computing*, 18(2):458–472, 2019. doi: 10.1109/TMC.2018.2836914.

Kesong Wu, Xianbin Cao, Peng Yang, Zongyang Yu, Dapeng Oliver Wu, and Tony Q. S. Quek. Qoe-driven video transmission: Energy-efficient multi-uav network optimization. *IEEE Transactions on Network Science and Engineering*, 11(1):366–379, 2024b. doi: 10.1109/TNSE.2023.3298782.

Manjia Wu, Weige Xie, Xiufang Shi, Panyu Shao, and Zhiguo Shi. Real-time drone detection using deep learning approach. In Limin Meng and Yan Zhang, editors, *Machine Learning and Intelligent Communications*, pages 22–32, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00557-3.

Liang Xiao, Yuzhen Ding, Jinhao Huang, Sicong Liu, Yuliang Tang, and Huaiyu Dai. Uav anti-jamming video transmissions with qoe guarantee: A reinforcement learning-based approach. *IEEE Transactions on Communications*, 69(9):5933–5947, 2021. doi: 10.1109/TCOMM.2021.3087787.

Xuedou Xiao, Wei Wang, Taobin Chen, Yang Cao, Tao Jiang, and Qian Zhang. Sensor-augmented neural adaptive bitrate video streaming on uavs. *IEEE Transactions on Multimedia*, 22(6):1567–1576, 2020. doi: 10.1109/TMM.2019.2945167.

Shimin Xiong, Bin Li, and Shiao Zhu. Dcgnn: a single-stage 3d object detection network based on density clustering and graph neural network. *Complex & Intelligent Systems*, 9(3):3399–3408, 2023. ISSN 2198-6053. doi: 10.1007/s40747-022-00926-z. URL https://doi.org/10.1007/s40747-022-00926-z.

Jiaqi Xu, Haipeng Yao, Ru Zhang, Tianle Mai, and Mohsen Guizani. Low latency and accuracy-guaranteed dnn inference for uav-assisted iot networks. *IEEE Transactions on Cognitive Communications and Networking*, pages 1–1, 2025. doi: 10.1109/TCCN.2025.3542443.

Junwei Xu, Bo Mo, Jie Zhao, Chunbo Zhao, Yimeng Tao, and Shuo Han. Cross-modal adaptive fusion object detection based on illumination-awareness. In *2024 39th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pages 931–938, 2024. doi: 10.1109/YAC63405.2024.10598791.

Renjie Xu, Saiedeh Razavi, and Rong Zheng. Edge video analytics: A survey on applications, systems and enabling techniques. *IEEE Communications Surveys & Tutorials*, 25(4):2951–2982, 2023. doi: 10.1109/COMST.2023.3323091.

Lei Yan, Qun Hao, Jie Cao, Rizvi Saad, Kun Li, Zhengang Yan, and Zhimin Wu. Infrared and visible image fusion via octave gaussian pyramid framework. *Scientific Reports, Nature*, 11, 2021. ISSN 2045-2322. doi: https://doi.org/10.1038/s41598-020-80189-1.

Guoyu Yang, Jie Lei, Zhikuan Zhu, Siyu Cheng, Zunlei Feng, and Ronghua Liang. Afpn: Asymptotic feature pyramid network for object detection. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2184–2189, 2023. doi: 10.1109/SMC53992.2023.10394415.

Kichang Yang, Juheon Yi, Kyungjin Lee, and Youngki Lee. Flexpatch: Fast and accurate object detection for on-device high-resolution live video analytics. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, pages 1898–1907, 2022. doi: 10.1109/INFOCOM48880.2022.9796984.

Abid Yaqoob, Zhenhui Yuan, and Gabriel-Miro Muntean. A uav-centric improved soft actor-critic algorithm for qoe-focused aerial video streaming. *IEEE Transactions on Vehicular Technology*, 73(9):13498–13512, 2024. doi: 10.1109/TVT.2024.3396349.

Ming Yin, Wei Liu, Xia Zhao, Yanjun Yin, and Yu Guo. A novel image fusion algorithm based on nonsubsampled shearlet transform. *Optik*, 125(10):2274–2282, 2014. ISSN 0030-4026. doi: https://doi.org/10.1016/j.ijleo.2013.10.064.

Ming Yin, Xiaoning Liu, Yu Liu, and Xun Chen. Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain. *IEEE Transactions on Instrumentation and Measurement*, 68(1):49–64, 2019. doi: 10.1109/TIM.2018.2838778.

Pingping Yu, Feng Chen, Jun Wang, Pingping Chen, Jiling Cai, and Mei Yang. Adaptive anti-packet loss strategy for real-time video streaming. In *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*, pages 2063–2068, 2022. doi: 10.1109/ICCC56324.2022.10065636.

Won Joon Yun, Byungju Lim, Soyi Jung, Young-Chai Ko, Jihong Park, Joongheon Kim, and Mehdi Bennis. Attention-based reinforcement learning for real-time uav semantic communication, 2021. URL https://arxiv.org/abs/2105.10716.

Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Muller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

Hongtao Zhang, Li Tan, Yuzhao Liu, Tianli Yuan, Haixia Zhao, and He Liu. Energy-efficient multi-uav collaborative path planning using levy flight and improved gray wolf optimization. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2024a. doi: 10.1109/IJCNN60899.2024.10650411.

Jiali Zhang, Zuriahati Mohd Yunos, and Habibollah Haron. Interactivity recognition graph neural network (ir-gnn) model for improving human–object interaction detection. *Electronics*, 12(2), 2023a. ISSN 2079-9292. doi: 10.3390/electronics12020470. URL https://www.mdpi.com/2079-9292/12/2/470.

Jialin Zhang, Xiang Huang, Jingao Xu, Yue Wu, Qiang Ma, Xin Miao, Li Zhang, Pengpeng Chen, and Zheng Yang. Edge assisted real-time instance segmentation on mobile devices. In *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*, pages 537–547, 2022a. doi: 10.1109/ICDCS54860.2022.00058.

Junbo Zhang and Kaisheng Ma. Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16650–16659, June 2022.

Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5126–5136, 2019. doi: 10.1109/ICCV.2019.00523.

Shi-Xue Zhang, Xiaobin Zhu, Jie-Bo Hou, and Xu-Cheng Yin. Graph fusion network for multi-oriented object detection. *Applied Intelligence*, 53(2):2280–2294, 2023b. ISSN 1573-7497. doi: 10.1007/s10489-022-03396-5. URL https://doi.org/10.1007/s10489-022-03396-5.

Xindi Zhang and Krishna Chandramouli. Critical infrastructure security against drone attacks using visual analytics. In Dimitrios Tzovaras, Dimitrios Giakoumis, Markus Vincze, and Antonis Argyros, editors, *Computer Vision Systems*, pages 713–722, Cham, 2019. Springer International Publishing. ISBN 978-3-030-34995-0.

Zeyao Zhang, Wen Dong, Jinpeng Bai, Chenguang Wang, Jiajin Qiu, Dilong Sun, and Xin Yang. Uav perspective small object detection with rgb-ir fusion perception. *Journal of Shanghai Jiaotong University (Science)*, July 2025. ISSN 1995-8188. doi: 10.1007/s12204-025-2835-3. URL https://doi.org/10.1007/s12204-025-2835-3.

Zhengning Zhang, Lin Zhang, Yue Wang, Pengming Feng, and Baochen Sun. Bidirectional parallel feature pyramid network for object detection. *IEEE Access*, 10:49422–49432, 2022b. doi: 10.1109/ACCESS.2022.3173732.

Zhiyong Zhang, Huaizu Jiang, and Hanumant Singh. Neuflow: Real-time, high-accuracy optical flow estimation on robots using edge devices. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5055, 2024b. doi: 10.1109/IROS58592.2024.10802353.

Gangming Zhao, Weifeng Ge, and Yizhou Yu. Graphfpn: Graph feature pyramid network for object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2743–2752. IEEE, 2021.

Tantan Zhao, Lijun He, Xinyu Huang, and Fan Li. Drl-based secure video offloading in mec-enabled iot networks. *IEEE Internet of Things Journal*, 9(19):18710–18724, 2022. doi: 10.1109/JIOT.2022.3161680.

Tantan Zhao, Fan Li, and Lijun He. Secure video offloading in multi-uav-enabled mec networks: A deep reinforcement learning approach. *IEEE Internet of Things Journal*, 11(2):2950–2963, 2024a. doi: 10.1109/JIOT.2023.3296613.

Tianyi Zhao, Maoxun Yuan, Feng Jiang, Nan Wang, and Xingxing Wei. Removal then selection: A coarse-to-fine fusion perspective for rgb-infrared object detection, 2024b. URL https://arxiv.org/abs/2401.10731.

Yongqiang Zhao, Rui Han, and Yuan Rao. A new feature pyramid network for object detection. In *2019 International Conference on Virtual Reality and Intelligent Systems (ICVRIS)*, pages 428–431, 2019. doi: 10.1109/ICVRIS.2019.00110.

Zhicheng Zhao, Wei Zhang, Yun Xiao, Chenglong Li, and Jin Tang. Reflectance-guided progressive feature alignment network for all-day uav object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025. doi: 10.1109/TGRS.2025.3574963.

Yuanhong Zhong, Xia Chen, Yongting Hu, Panliang Tang, and Fan Ren. Bidirectional spatio-temporal feature learning with multiscale evaluation for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8285–8296, 2022. doi: 10.1109/TCSVT.2022.3190539.

Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 843–851, 2019. doi: 10.1109/ICCV.2019.00093.

Junqing Zhu, Yuxuan Wu, and Tao Ma. Multi-object detection for daily road maintenance inspection with uav based on improved yolov8. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):16548–16560, 2024. doi: 10.1109/TITS.2024.3437770.

Linxiang Zhu, Feifei Lee, Jiawei Cai, Hongliu Yu, and Qiu Chen. An improved feature pyramid network for object detection. *Neurocomputing*, 483:127–139, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.02.016. URL https://www.sciencedirect.com/science/article/pii/S0925231222001588.